12-2016

Visual analytics of location-based social networks for decision support

Junghoon Chae
Purdue University

VISUAL ANALYTICS OF LOCATION-BASED SOCIAL NETWORKS FOR DECISION SUPPORT

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Junghoon Chae

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2016

Purdue University

West Lafayette, Indiana

Graduate School Form 30
(Revised 08/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Junghoon Chae

Entitled: Visual Analytics of Location-based Social Networks for Decision Support

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

DAVID S. EBERT

JI SOO YI

NIKLAS ELMQVIST

YUN JANG

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

DAVID S. EBERT

Approved by Major Professor(s): _____

_____

Approved by: V. Balakrishnan 12/05/2016

Head of the Department Graduate Program Date

This thesis is dedicated to my parents and family.

ACKNOWLEDGMENTS

I would like to express my deep gratitude to my advisor, Dr. David S. Ebert, for supporting and guiding me through to the completion of this thesis. I am also grateful to Dr. Yun Jang. He spent a considerable amount of time to provide sound advice and feedback to my research during the past few years. I would also like to express my sincere gratitude to my friend and mentors, Insoo and SungYe, whose support and advice have been invaluable. I would like to acknowledge help from all colleagues at the VACCINE Center, particularly Abish and Jiawei for their support throughout the years. Finally, I am greatly indebted to my wife, Jiyoung. I could never have done this without you, my forever sweetheart.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES viii

ABBREVIATIONS xii

ABSTRACT xiii

1 INTRODUCTION 1

1.1 Visual Analytics of Location-based Social Networks for Abnormal Event Detection 2

1.2 Visual Analytics for Public Behavior Analysis in Disaster Events 3

1.3 Visual Analytics of Anomalous Human Movement Analysis 4

1.4 Visual Analytics of Forecasting the Flow of Human Crowds 5

1.5 Thesis Statement 7

2 BACKGROUND AND RELATED WORK 9

2.1 Visual Analytics of Movement Data 9

2.2 Visual Analytics of Location-based Social Networks 11

2.3 Event Detection and Topic Analysis of Social Media 12

2.4 Disaster Management based on Social Media Analysis 14

2.5 Human Movement Analysis using Location-based Social Networks . . . 15

2.6 Predicting Movement 16

3 VISUAL ANALYTICS OF SPATIOTEMPORAL SOCIAL MEDIA FOR ABNORMAL EVENT DETECTION 19

3.1 Spatiotemporal Social Media Analytics for Event Examination 20

3.1.1 Topic Extraction 21

3.1.2 Abnormality Estimation using Seasonal-Trend Decomposition . 24

3.1.3 Detection Model 25

Page

3.2 Interactive Analysis Process . . . 26
3.2.1 Social Media Retrieval and Analysis System . . . 26
3.2.2 Visual Topic Exploration and Event Evaluation . . . 28
3.3 Case Study . . . 30
3.3.1 Ohio High School Shooting . . . 31
3.3.2 Occupy Wall Sreet . . . 32
3.3.3 2011 Virginia Earthquake . . . 34
3.4 Discussion . . . 36
3.5 Summary . . . 38
4 VISUAL ANALYTICS OF MICROBLOG DATA FOR PUBLIC BEHAVIOR ANALYSIS IN DISASTER EVENTS . . . 39
4.1 Problem Statement and Interactive Analysis Process Design . . . 40
4.2 Spatiotemporal Analysis . . . 42
4.2.1 Spatial Analysis . . . 42
4.2.2 Spatial Decision Support . . . 45
4.2.3 Temporal Pattern Analysis . . . 51
4.2.4 Spatiotemporal Visualization . . . 53
4.3 Discussion and Evaluation . . . 54
4.4 Summary . . . 55
5 TRAJECTORY-BASED VISUAL ANALYTICS FOR ANOMALOUS HUMAN MOVEMENT ANALYSIS . . . 57
5.1 System Overview . . . 58
5.1.1 Trajectory Extraction . . . 59
5.1.2 Data Analysis . . . 60
5.1.3 Visualization and Analysis . . . 63
5.2 Improving Analysis using Multi-Context Information . . . 66
5.2.1 Keyword Extraction and Visualization . . . 67
5.2.2 Additional Context Information . . . 70
5.3 Case Study . . . 71

Page

5.3.1 Boston Marathon Explosion 72
5.3.2 Purdue University Shooting 72
5.4 Summary 73
6 FORECASTING THE FLOW OF HUMAN CROWDS 74
6.1 Flow Data Modeling, Forecasting, and Visualization 75
6.1.1 Directional Density Estimation: Classification of Fixed Direction Sector 75
6.1.2 Directional Density Estimation: Considering Road Direction . . 77
6.1.3 Flow Smoothing 78
6.1.4 Forecasting Future Flow 81
6.1.5 Missing-Data Imputation 85
6.1.6 Visualization of Multi-Vector Fields 85
6.2 Evaluation 89
6.2.1 Spatial Error Analysis 93
6.3 Discussion 96
6.3.1 Grid Size 97
6.3.2 Regional Differentials in Forecasting Error Rate 98
6.4 Summary 98
7 CONCLUSIONS 99
7.1 Future Work 100
LIST OF REFERENCES 102
VITA 113

LIST OF TABLES

Table Page

3.1 An example of extracted topics and their proportions. We extracted topics from Tweets written on August 23, 2011 around Virginia, where an earthquake occurred on this day. One can see that topics consisting of ordinary and unspecific words can have high proportion values, while the earthquake related topics have a relatively low proportion value. 21

3.2 An example of topic model results depending on the number of iteration steps in the LDA process. The topics are extracted from the Tweets posted in New York City on September 17 and 18, 2011 where the Occupy Wall Street protest movement began and a famous festival, *San Gennaro* occurred. A higher number of sampling iterations provides a better topic retrieval describing the two different events. 23

LIST OF FIGURES

Figure Page

3.1 Overview of our iterative analysis scheme for event detection and examination. 27

3.2 Examining the location of the Chardon high school shooting with a text aggregating content lens. 30

3.3 Social media analysis system including message plots on a map, abnormality estimation charts and tables for message content and topic exploration. It can be seen, how the Ohio High School Shooting on February 27, 2012 is examined using the system. The selected messages, marked as white dots on the map, show retrieved Tweets that are related to the event. 32

3.4 Cross validation of an event using Twitter, Flickr, and YouTube data for the Occupy Wall Street Protests. The protests occurred on Sep. 17 and 30, Oct. 5 and 15. The line charts show the remainder components R (blue) and the original data volumes Y (red) for the STL evaluation. The scales on the right and left side of each chart view are adapted to the maximum values. 33

3.5 Abnormality and correlation on multiple social media sources. As a result of high z-scores around the same time periods, we found a strong correlation between the three social media sources. Marked regions correspond to periods where at least 2 providers received scores over 2.0. 34

3.6 Virginia earthquake on August 23rd, 2011. Our abnormal event detection system detects the earthquake event using our STL based anomaly detection algorithm. The abnormality degree is extremely high on August 23rd, 2011 (times are given in UTC). 35

4.1 Overview of our interactive analysis scheme for public behavior analysis using social media data. 41

4.2 Spatial user-based Tweet distribution in the Manhattan area in New York City during four hours right after the evacuation order (from 12:00 PM to 4:00 PM on October 28th, 2012 (Right)). Previous distribution of Tweets on 14th (Left) and 21st (Center). 43

4.3 Twitter user distribution on the eastern coast area in New Jersey, after the hurricane passed over the area on October 31st (Right). Previous distribution on October 24th is shown on the Left. 44

Figure Page

4.4 Distribution of Twitter users of each consecutive date (Oct. 26 $\sim$ 30, 2012), who post hurricane related Tweets on the southeastern (1 and 2) and northeastern coast (3, 4, and 5) area of the United States. We can see the variance of Twitter user reactions along the track of the hurricane center locations. . . . 47

4.5 Spatial pattern of Twitter users during 24 hours in the city of Moore after damages from a strong tornado. Relatively many people moved to severely damaged areas after the disaster. This situation is much different from the previous normal situation (1). We selected a specific region (2) that includes severely damaged areas in order to extract topics (3) from Tweets within the selected area. 49

4.6 Topic cloud: Topics from Tweets within the selected area in Figure 4.5 (2) are ordered by their abnormality scores. 49

4.7 Abnormality of the first topic in Figure 4.6. The abnormality score of the topic had significantly increased when the tornado hit the region on May 20th (Marked region). 51

4.8 Temporal analysis for public behaviors during the disaster event, Sandy. Top shows our entire system view. The bar chart (Bottom) for the number of Twitter users within the selected region including a supermarket in Figure 4.2 (Right) in four hour intervals is shown. We see that many people went to the supermarket right after the evacuation order. 52

4.9 Visualization for spatiotemporal social media data (Left). A hexagon represents the spatial (position) and temporal (color) information of a Tweet. Hurricane evacuation map [125] (Right). Residents in Zone A (red) faced the highest risk of flooding, Zone B (yellow) and Zone C (green) are moderate and low respectively. 53

5.1 Overview of our iterative analysis scheme for human common movement discovery and anomaly analysis. 58

5.2 Supplementing a sparse trajectory (Blue) using route direction information (Yellow). 59

5.3 Discovering a common sub-trajectory. 61

5.4 Similarity measurement for two line segments. 62

5.5 The process of discovering common human movement patterns using location-based social networks data. Visualization of sub-trajectory clusters (right). The thickness of each trajectory represents the size of the cluster. 65

5.6 Visualization of sub-trajectory clusters. The thickness of each trajectory represents the size of the cluster. 66

Figure Page

5.7 Clustering results depending on two parameters: ε and *MinLns*. Top ($\varepsilon = 25$, $MinLns = 3$) is optimal. Center ($\varepsilon = 25$, $MinLns = 4$) shows less number of trajectory clusters. Bottom ($\varepsilon = 30$, $MinLns = 2$) shows more. 67

5.8 The trajectories (red and orange) shows the human movement patterns around the finish line at the Boston Marathon 2013 during 2 hours after the explosions. The trajectories (blue) represent the movements for the normal situation (the same time period of the same event in 2014). The two markers indicate the locations of the two explosions. 68

5.9 The extracted keywords along the trajectory close to the explosion locations show a strong relationship to the explosions (top). The chronologically displayed photos (bottom) extracted from the same trajectory show the scenes of evolving situations. 69

5.10 The trajectories (red and orange) shows the human movements around the campus during 2 hours after the shooting. The normal trajectories (blue) extracted from the same time period on normal day. Photos (1), News reports (2), Keywords (3), and Webcam videos (4). The green rectangles indicate the locations of the web cameras around the campus. The yellow one is the selected camera. The marker indicate the building where the accident occurred. 71

6.1 Trajectory tessellation and directional density estimation 75

6.2 Trajectory tessellation and directional density estimation based on shortest path considering road directions 78

6.3 Directional density of New York City resulted by the method 1 (Left) and the method 2 (Right) 79

6.4 Directional density of New York City resulted by the method 1 (Left) and the method 2 (Right) 80

6.5 Each space's directional density is computed based on its neighbors' and global trends. 81

6.6 Local directional density (purple arrows) and global (orange arrow) trends of taxi movements in southern Manhattan between 6:00 AM and 10:00 AM. . . 82

6.7 An example of the smoothing result. Sub-space B does not have density due to sparseness data (Left). Sub-space has the density (Right). 83

6.8 The process of multi-vector field prediction. 83

6.9 Magnitude values of a specific direction in a grid cell for 150 days with 4 hours (Top). Interpolated result (Bottom). The blue circles are the observed data points and the red ones are imputed ones. 86

Figure Page

6.10 The flows of Taxi data between 7:00 AM and 9:00 AM. Multi-vector fields representing the directional densities (Top). Particle advection (Left-Bottom). The paths of long-life particles (Right-Bottom). The red color indicates a high probability, whereas, the yellow represents a low probability. 87

6.11 Comparison of forecasting results by two different approaches. 90

6.12 Error rates for varying grid sizes for the Twitter and the taxi datasets. 91

6.13 Error rates for varying grid sizes for the Twitter data under the different combinations of the methods. R is the method considering road conditions and I is the missing-data imputation method. 92

6.14 Error rates for varying grid sizes for the taxi data under the different approach conditions. 93

6.15 Visual comparison of particle advection: Taxi in Porto. (Top: Observed data, Bottom: Forecasting data) 94

6.16 Visual comparison of paths of long-life particles: Taxi in Porto. (Top: Observed data, Bottom: Forecasting data) 95

6.17 Visual comparison of particle advection: Twitter in New York City. (Left: Observed data, Right: Forecasting data) 96

6.18 Visual comparison of paths of long-life particles: Twitter in New York City. (Left: Observed data, Right: Forecasting data) 96

6.19 Spatial error analysis: Regional differentials in error rate. Red color indicates the high error rate; Green color indicates the low error rate. 97

ABBREVIATIONS

LBSN	Location-Based Social Network
LDA	Latent Dirichlet Allocation
STL	Seasonal-Trend Decomposition procedure based on Loess smoothing
OD	Origin-Destination
RBF	Radial Basis function
HMM	Hidden Markov Models
DHR	Dynamic Harmonics Regression
SARIMA	Seasonal AutoRegressive Intergrated Moving Average
DBSCAN	Density-based spatial clustering of applications with noise
GPS	Global Positioning System
RMSE	Root Mean Square Error
NRMSE	Normalized Root Mean Square Error

ABSTRACT

Chae, Junghoon Ph.D., Purdue University, December 2016. Visual Analytics of Location-based Social Networks for Decision Support. Major Professor: David S. Ebert.

Recent advances in technology have enabled people to add location information to social networks called Location-Based Social Networks (LBSNs) where people share their communication and whereabouts not only in their daily lives, but also during abnormal situations, such as crisis events. However, since the volume of the data exceeds the boundaries of human analytical capabilities, it is almost impossible to perform a straightforward qualitative analysis of the data. The emerging field of visual analytics has been introduced to tackle such challenges by integrating the approaches from statistical data analysis and human computer interaction into highly interactive visual environments.

Based on the idea of visual analytics, this research contributes the techniques of knowledge discovery in social media data for providing comprehensive situational awareness. We extract valuable hidden information from the huge volume of unstructured social media data and model the extracted information for visualizing meaningful information along with user-centered interactive interfaces. We develop visual analytics techniques and systems for spatial decision support through coupling modeling of spatiotemporal social media data, with scalable and interactive visual environments. These systems allow analysts to detect and examine abnormal events within social media data by integrating automated analytical techniques and visual methods. We provide comprehensive analysis of public behavior response in disaster events through exploring and examining the spatial and temporal distribution of LBSNs. We also propose a trajectory-based visual analytics of LBSNs for anomalous human movement analysis during crises by incorporating a novel classification technique. Finally, we introduce a visual analytics approach for forecasting the overall flow of human crowds.

1. INTRODUCTION

Since the high global Internet penetration rate and the Web 2.0 era, humans have become a biggest data source. Humans extremely fast generate a variety of big data using multiple devices, such as personal computers, smartphones and tablets in multiple environments, such as social networks and (micro)blogs. The data generated by humans is worth understanding, estimating, and predicting their behavior in many areas including marketing, research, and public administration and management. Also recent advances in technology have enabled people to add location information to social networks called Location-Based Social Networks (LBSNs) where millions of people share their communication and whereabouts not only in their daily lives, but also during abnormal situations, such as crisis events. Such spatiotemporal data not only provides location-embedded information, but also bring new solutions to a wide range of challenges in analyzing social behaviors and interaction in the physical world.

However, the data has challenging issues. Since the volume of the data exceeds the boundaries of human analytical capabilities and normal computing performance, it is almost impossible to perform a straightforward qualitative analysis of the data. Also, the contents of the data are usually unstructured and have a high degree of noise. To address these challenges, researchers have proposed visual analytics that is defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [1]. Currently, many visual analytics techniques integrating approaches from data mining, statistics, and human computer interaction have been proposed to combine the computing power of machines and human analytical capabilities.

In this thesis, We propose four visual analytics approaches that provide users with scalable and interactive visual spatiotemporal social media data analysis supporting comprehensive situational awareness for spatial decision support. In this chapter, the following

four sections describe an overview of each approach and Section 1.5 provides our thesis statement.

1.1 Visual Analytics of Location-based Social Networks for Abnormal Event Detection

Internet users from all over the world have created a large volume of time-stamped, geo-located data. Such spatiotemporal data has immense value for increasing situational awareness of local events, providing insights for investigations and understanding the extent of incidents, their severity, and consequences, as well as their time-evolving nature. In analyzing social media data, researchers have mainly focused on finding temporal trends according to volume-based importance. Thus, a relatively small volume of relevant messages for situational awareness are usually buried by a majority of irrelevant data. Finding and examining these messages without smart aggregation, automated text analysis and advanced filtering strategies is almost impossible and extracting meaningful information is even more challenging.

In this thesis, we present a visual analytics approach that provides users with scalable and interactive social media data analysis and visualization including the exploration and examination of abnormal topics and events within various social media data sources, such as Twitter, Flickr and YouTube. In order to find and understand abnormal events, the analyst can first extract major topics from a set of selected messages and rank them probabilistically using Latent Dirichlet Allocation (LDA) [2], which extracts and probabilistically ranks major topics contained in textual parts of the social media data. The ranks of the categorized topics generally provide a volume-based importance, but this importance does not reflect the abnormality or criticality of the topic. In order to obtain a ranking suitable for situational awareness tasks, we discard daily chatter by employing a Seasonal-Trend Decomposition procedure based on Loess smoothing (STL) [3]. Our whole analysis process, including the application of automated tools, is guided and informed by an analyst using a highly interactive visual analytics environment. It provides tight integration

of semi-automated text-analysis and probabilistic event detection tools together with traditional zooming, filtering and exploration following the Information-Seeking Mantra [4].

1.2 Visual Analytics for Public Behavior Analysis in Disaster Events

For emergency and disaster management, analysis of public behavior, such as how people prepare and respond to disasters, is important for evacuation planning. As social media has played a pervasive role in the way people think, act, and react to the world, in even emergency situations, people seek social confirmation before acting in response to a situation, where they interact with others to confirm information and develop a better informed view of the risk [5]. Moreover, a growing number of people are using LBSN services, such as microblogs, where they create time-stamped, geo-located data and share this information about their immediate surroundings using smart phones with GPS. Such spatiotemporal data has great potential for enhancing situational awareness during crisis situations and providing insight into the evolving event, the public response, and potential courses of action.

However, finding meaningful information from social media is challenging because the large volume of unstructured social media data hinders exploration and examination. Even though we could extract certain information from the data, it is not always easy to determine whether the analysis result of the extracted information is meaningful and helpful. Thus, there is a need for advanced tools to handle such big data and aid in examining the results in order to understand situations and glean investigative insights. Given the incomplete, complex, context-dependent information, a human in this analysis and decision-making loop is crucial. Therefore, a visual analytics approach offers great potential through interactive, scalable, and verifiable techniques, helping analysts to extract, isolate, and examine the results interactively.

In this research, we present an interactive visual analytics approach for spatiotemporal microblog data analysis to improve emergency management, disaster preparedness, and evacuation planning. We demonstrate the ability to identify spatiotemporal differences in

patterns between emergency and normal situations, and analyze spatial relationships among spatial distributions of microblog users, locations of multiple types of infrastructure, and severe weather conditions. Furthermore, we show how both spatiotemporal microblog and disaster event data can help the analysts to understand and examine emergent situations.

1.3 Visual Analytics of Anomalous Human Movement Analysis

Analysis of human movement patterns is important for urban planning [6], traffic forecasting [7], and understanding the pandemic spread of diseases [8]. For crisis and disaster events, movement analysis, such as where people move to/from and how people respond to disasters, is also critical for evacuation management. Unfortunately, finding meaningful data is challenging and collecting relevant data can be costly. However, the rapid development and increasing availability of mobile communication and location acquisition devices allow people to share location data using existing social networks. These location-based social networks (LBSNs) have been gaining attention as promising data sources for analyzing human movements. Particularly, trajectories—sequences of geo-referenced data nodes of each user—extracted from such LBSNs provide opportunities and solutions to challenges in human movement analysis [9–11]. In addition, semantic context of the data enhances understanding of local events and human movements [12, 13].

Previous studies have mainly focused on finding regular movement patterns using spatial data. They have demonstrated that human movements are normally influenced by geographic constraints, life patterns, and spatial and temporal events, such as local festivals and holiday seasons [14, 15]. However, during disaster events, since human movement patterns (e.g., volume and direction of movements) are unusual compared to normal situations, a new approach is required to analyze the movements. Also, analyzing location data alone has shown limitations in achieving situational awareness of local events. For example, they cannot answer why people move and what situations occur.

To address these challenges, we propose a trajectory-based visual analytics system for anomalous human movement analysis during disasters using multi-type online media. Our

system extracts geo-location information of each data node from LBSNs and generates trajectories using the information. The generated raw trajectories, however, do not have enough fine-grained spatial positions. We supplement the sparse positions in the trajectories using route information between each position. We group the individual trajectories into classes of similar sub-trajectories using a trajectory clustering model based on the partition-and-group framework [16]. This enables users to discover sub-common patterns, rather than finding common patterns as a whole. We also propose a classification model based on historical data for detecting abnormal movements using human expert interaction. In addition, we integrate multiple visual representations using relevant context extracted from different online media sources, such as Tweet text, shared photos, public webcam videos, and news media to allow users to discover and analyze anomalous human movement patterns; thereby, improving situational awareness in disaster management situations.

1.4 Visual Analytics of Forecasting the Flow of Human Crowds

Forecasting human crowd flows plays a significant role in a range of applications, from urban and traffic planning [6, 7] to predicting epidemic dynamics [8, 17]. Various location-based services are also highly dependent on foreseeing human movement patterns [18]. The volume and variety of data that capture the different aspects of human mobility have greatly increased due to ubiquitous crowdsourced activities and the advent of several location-based social networking services. The potential impacts and availability of such data have caught the attention of researchers from various domains. They have put considerable efforts into understanding and predicting human mobility patterns [19–21]. They have also discovered that human mobility behaviors can contain discernible patterns that can be used for forecasting purposes. For example, Song et al. [22] reported a 93% potential predictability in human mobility from anonymized mobile phone data, and although the overall travel patterns were vastly different, the variability in predictability was found to be significantly low.

Many previous techniques for predicting human movements have treated the individual mobility behaviors as discrete entities and focused on predicting the next destinations of the individuals based on the observations of their past movement patterns or frequent behaviors of similar users [20,21,23]. While recent work has made progress in predicting human mobility patterns, the proposed models that are based on movements of individuals suffer from many limitations. For example, the data modeling techniques used in these methods (e.g., Hidden Markov Model) typically require extensive training of the data models using historical observed data [20,24]. This process can be expensive and rate limiting, especially in case of large scale datasets (e.g., taxi data in large urban regions), and can further severely restrict interactive visual analytic system behaviors. Other challenges include privacy concerns for the individuals [20, 25]. Furthermore, the individual spatial sequence trajectory datasets, especially those derived from location-based social networks, can also suffer from data sparsity and noise problems [26, 27]. These can often prove to be prohibitive for accurate data modeling. Standard methods to mitigate for these challenges include clustering techniques that summarize the overall movement paths in trajectory data analysis. These require analysts to carefully select appropriate abstraction levels for clustering in order to prevent the original vector flow data from getting distorted. However, when these abstraction levels are carefully chosen, the analysis of collective flows of human crowds can provide new insights that may not be available at finer granularity levels [28, 29].

To this end, this thesis presents a space-based approach for forecasting the flow of human crowds. Our work is motivated by weather simulation and forecasting modeling techniques that are built using local atmospheric observations from weather stations. In this work, we embed individual movements into a two-dimensional Euclidean space and model for the space instead of the moving objects. In other words, given a space with a large number of moving objects, we discretize the space into smaller sub-spaces and model the movement flows for each sub-space. Our model forecasts the future flow based on observed historical patterns of each sub-space using a seasonal trend analysis technique [3]. We then combine the results to visualize the future flow as a whole for the entire space.

Our approach consists of a directional flow density estimation method that preserves the original paths and directions of moving entities, and a flow smoothing method based on both local and global trajectory trends to mitigate for the data sparseness and noise issues. We also provide a novel visualization technique for showing the probability density distribution of flow. We demonstrate our work using location-based social media data and GPS tracking human and taxi data.

1.5 Thesis Statement

This research contributes the techniques of knowledge discovery in social media data to provide comprehensive situational awareness for decision making. We extract valuable hidden information from the huge volume of unstructured social media data and model the extracted information for visualizing meaningful information along with user-centered interactive interfaces. This thesis presents the design and development of visual analytics techniques and systems for spatial decision support through coupling modeling of spatiotemporal social media data, with scalable and interactive visual environments. The major contributions of this work are the following:

1. Abnormal topic detection within social media data by combining the STL and the LDA topic model

2. Design of visual analytics system that enables integration of LBSN data with geospatial disaster and infrastructure data for supporting spatial decision-making in crisis management

3. Common human movement pattern discovery from LBSNs using a trajectory clustering model based on the partition-and-group framework

4. Abnormal human mobility pattern detection and visualization using a trajectory-based anomaly detection model

5. Development of visual means to improve human movement analysis using semantic context available from multiple online media sources

6. Development of a new method to estimate the directional flow density that represents the overall movement direction and preserves the original paths and directions of moving entities

7. Development of a new flow smoothing method based on both local and global movement trends for improving forecast accuracy by mitigate the effect of the data sparsity and noise

8. Development of a new model to forecast vector field data with the STL by transforming data to a series of magnitude values from the smoothed representative vectors

2. BACKGROUND AND RELATED WORK

In recent years social media data has become a popular topic in a range of application domains. Researchers in the fields of data mining and visual analytics have found through studies among users and domain experts, that the analysis of such data can be essential for spatiotemporal situational awareness [30, 31]. Also, several researchers have proposed and presented systems for social media analysis and important studies covering the use of social media during crisis events have been conducted. Thus, as the size of social media data increases, scalable computational tools for the effective analysis and discovery of critical information within the data are a vital research topic. This section presents previous work that has focused on visual analytics of movement data and LBSNs, crisis related social media exploration, visualization, and human movement analysis using LBSNs, and predictive movement data exploration.

2.1 Visual Analytics of Movement Data

With the belief that location-based visual analysis can intuitively assist users to understand environments and find out key events [32–34], much research has been performed that led to the development of various techniques and approaches in the area of visual analysis of trajectory-based movement [35]. Here, we describe the analysis approaches according to Andrienko's categorization [36] (direct depiction, summarization, and pattern extraction).

The techniques in the direct depiction category directly visualize data on a screen and allow analysts to extract information with interaction methods. In general, visualizations of this type extensively use a line-based movement representation such as visualizations with polyline paths [37], stacking-based attributed trajectories [38], arrows [39] and space time cubes [40]. Even thought the techniques in this category are intuitive to understand,

they have a visual clutter issue when trajectories are very complex, such as aircraft trajectories [41]. This creates a need for a trajectory summarization approach.

In the summary visualization techniques, statistical computations are performed on the data and then the results of the computations are visualized. One of the popular computation techniques is aggregation\clustering [42, 43]. In particular, aggregation helps users to understand sparse trajectories by reducing uncertainties in terms of time and space [32]. Often, the computation results generate additional characteristics that permit a new analysis in different aspects. For example, Andrienko et al.'s algorithm transforms geo-located data into individual-group relation data that allows another type of movement analysis [44]. This allows researchers to design new visualization techniques or reuse many advanced visualization techniques for movement data analysis including multivariate glyph visualization [45], density map [46], origin-destination (OD) map [47, 48], and Flowstrates [49]. Some of previous work also adopt vector field related approaches [42, 50, 51]. Poco et al. [50] take account the road network as a graph and compute plausible routes for the New York taxi trips by computing possible paths and choose a path whose distance is similar enough to the actual distance. They compute the vector-valued traffic function based on the inferred paths and visualize the function using the particle advection techniques to show traffic mobility dynamics. Ferreira et al. [42] generate vector fields based on trajectory data and classify the groups of movements by clustering the vector fields. Nascimento et al. [51] also derive vector fields from trajectory data and represent the movement patterns as mixture of models.

The goal of the techniques in the pattern extraction category is to provide an environment for investigating hidden patterns in various aspects (e.g., themes [52], semantics [53], context [54], co-occurrence [55]) with additional complex computations based on algorithms from trajectory mining [56]. Trajectory mining itself is a wide research theme that includes most of the algorithms introduced in this section. In this work, we provide a visual predictive trajectory analytics environment based on a new flow forecasting model adapting a seasonal trend analysis. In addition, we provide a new visualization technique to show directional density of the flow of trajectories.

2.2 Visual Analytics of Location-based Social Networks

As social media platforms move towards LBSNs. researchers have proposed various approaches to analyze spatiotemporal document collections, in general, and spatiotemporal social media data, in particular. VisGets [57] provides linked visual filters for the space, time and tag dimensions to allow the exploration of datasets in a faceted way. The user is guided by weighted brushing and linking, which denotes the co-occurrences of attributes. Further works demonstrate the value of visualizing and analyzing the spatial context information of microblogs for social network users [58] or third parties like crime investigators [59] and urban planners [60]. With Senseplace2, MacEachren et al. [30] demonstrate a visualization system that denotes the message density of actual or textually inferred Twitter message locations. The messages are derived from a textual query and can then be filtered and sorted by space and time. Their work also has shown that social media can be a potential source for crisis management. With ScatterBlogs [61], our own group developed a scalable system enabling analysts to work on quantitative findings within a large set of geolocated microblog messages. In contrast to Senseplace2, where the analysts still have to find and manage the appropriate keywords and filters to gather relevant messages in the high volume of insignificant messages, we propose a semi-automatic approach that finds possibly relevant keywords and ranks them according to their ‘abnormality’.

Special LBSN for certain domains, like Bikely [62] and EveryTrail [63] have an even stronger focus on the sharing and tracing of user locations. Ying et al. [64] present various location based metrics using spatial information of these LBSNs to observe popular people who receive more attention and relationships within the network. Similarly, there are many related works for non-spatial temporal document collections, for example IN-SPIRE [65], which is a general purpose document analysis system that depicts document clusters on a visual landscape of topics.

2.3 Event Detection and Topic Analysis of Social Media

One of the major challenges in analyzing social media data is the discovery of critical information obscured by large volumes of random and unrelated daily chatter. Due to the nature of microblogging, message streams like Twitter are very noisy compared to other digital document collections. Recently, many researchers have tried to solve this challenge by means of automated and semi-automated detection and indication of relevant data.

Sakaki et al. [31] propose a natural disaster alert system using Twitter users as virtual sensors. In their work, they were able to calculate the epicenter of an earthquake by analyzing the delays of the first messages reporting the shock. Weng and Lee [66] address the challenge by constructing a signal for each word occurring in Twitter messages using wavelet analysis, thereby making it easy to detect bursts of word usage. Frequently recurring bursts can then be filtered by evaluating their auto-correlation. The remaining signals are cross correlated pairwise and clustered using a modularity-based graph partitioning of the resulting matrix. Due to the quadratic complexity of pairwise correlation, they rely on heavy preprocessing and filtering to reduce their test set to approx 8k words. As a result, they detected mainly, large sporting events, such as soccer world cup games, and elections. Our approach, in contrast, provides a set of topics through a probabilistic topic extraction algorithm which can be iteratively applied to subsets and subtopics within user selected message sets.

Lee and Sumiya [67] as well as Pozdnoukhov and Kaiser [68] present methods to detect unusual geo-social events by measuring the spatial and temporal regularity of Twitter streams. Lee and Sumiya propose a concept to detect unusual behavior by normalizing the Twitter usage in regions of interests which are defined by a clustering-based space partitioning. However, their results are mainly a measurements of unusual crowd behavior and do not provide further means for analyzing the situation. Pozdnoukhov and Kaiser observe abnormal patterns of topics using spatial information embedded in Twitter messages. Similar to our approach, they apply a probabilistic topic model (Online Latent Dirichlet Allocation) as a means of analyzing the document collection. A Gaussian RBF kernel den-

sity estimation examines the geo-spatial footprint of the resulting topics for regularities. The usual message count of identified areas is then learned by a Markov-modulated non-homogeneous Poisson process. The spatial patterns are shown as a static heat map. The resulting system does not provide interactive analytics capabilities.

Recently, researchers have applied LDA topic modeling to social media data to summarize and categorize Tweets [69] and find influential users [70]. Zhao et al. [69] demonstrate characteristics of Twitter by comparing the content of Tweets with a traditional news medium, such as the New York Times. They discuss and adapt a Twitter-LDA model and evaluate this model against the standard topic model and the so-called author-topic model [71], where a document is generated by aggregating multiple Tweets from a single user, in terms of meaningfulness and coherence of topics and Twitter messages. In this work, we do not use the author-topic model, since a users Tweet timeline is usually a heterogeneous mixture of unrelated comments and messages and not a homogenous framework of interrelated topics like a traditional document. Furthermore, the evaluation of Zhao et al. [69] shows that the standard model has quite reasonable topic modeling results on Tweets, although the Twitter-LDA model outperforms the standard model. Works from Ramage et al. [72] also show promising results in LDA based Twitter topic modeling by evaluating another type of LDA model (Labeled LDA) [73]. ParallelTopics [74] also extracts meaningful topics using LDA from a collection of documents. The visual analytics system allows users to interactively analyze temporal patterns of the multi-topic documents. The system, however, does not not deal with spatial information, but takes an abnormality estimation into account.

In our previous work [75], we proposed a spatiotemporal anomaly overview based on a streaming enabled clustering approach that is applied for each term in the dataset individually. The resulting clusters can be used to generate a spatially and temporally explorable term map of large amounts of microblog messages as an entry point for closer examination. Even though the scalable event detection and our current approach share the same workbench, they can be used independently as well as complementary. The combination of LDA and STL allows for an ad-hoc analysis of a user selected set of messages regard-

ing the topical distribution of messages and the abnormal presence of topics. Due to this characteristic, it provides an iterative analysis loop for qualitative analysis and drill down operations.

2.4 Disaster Management based on Social Media Analysis

Most recent analysis environments for crisis-related social media exploration and visualization are from MacEachren et al. [30], Marcus et al. [76], and Thom et al. [75]. Their systems combine traditional spatial and geographic visualizations with means for automated location discovery, trend and outlier search, anomaly and event discovery, large scale text aggregation and highly interactive geovisual exploration. Approaches putting less focus on visualizations and more on fully automated data mining mechanisms have been proposed by Sakaki et al. [31] that use Kalman and Particle Filters to detect the location of earthquakes and typhoons based on Twitter. Various techniques for spatiotemporal data analysis and anomaly detection using visualization or machine learning techniques have been proposed by Andrienko et al. [77], Lee and Sumiya [67], and Pozdnoukhov and Kaiser [68]. Twitcident from Abel et al. [78] provides a web-based framework to search and filter crisis-related Tweets. Using the Netherlands emergency broadcast system, Twitcident automatically reacts on reported incidents and collects related information from Twitter based on semantic enrichment. In all these system the focus is primarily on individual messages and aggregated message volumes and how insight can be generated by understanding their content. In contrast, our system investigates a more user focused approach that tries to identify the whereabouts and movements of people in order to understand mass behavior.

Researchers have also examined the usage of Twitter during incidents and disasters. Terpstra et al. [79] investigate more than 90k Twitter messages that were sent during and after a storm hit the Belgium *Pukkelpop* musicfestival in 2011. They categorize Tweets into warnings about the severe weather conditions, rumors and self organization of relief measures. They show that valuable information for crisis response and decision support can

be gathered from the messages. Vieweg et al. [80] investigate the differences in reaction to different crisis events. For their study they investigate eyewitness reports in Twitter from people that were affected by Oklahoma Grassfires in April 2009 and Red River Floods in March and April 2009. Their research also demonstrates the high value that the extraction of meaningful comments from crisis-related communication can have to generate insights. Furthermore, Heverin et al. [81] demonstrate that Twitter can also be a useful source of information for smaller events as they investigate the reaction to a shooting of four police officers and the subsequent search for the suspect that took place in the Seattle-Tacoma area. Based on the collection and categorization of 6000 messages they are able to show that citizens use the service to communicate and seek information related to the incident.

In this thesis, we also present case studies on crisis-related information gathered from Twitter data. However, in contrast to the discussed studies that harvest information directly out of the content of the messages, our method is primarily based on observing movement patterns and identifying local hotspots in order to learn about the effects of the crisis and the performance of evacuation measures.

2.5 Human Movement Analysis using Location-based Social Networks

As many social networks move towards LBSNs, researchers have proposed various approaches to analyze spatiotemporal social media data. Adrienko et al. [82] describe a visual analysis approach for exploring Tweet text and spatiotemporal patterns. Krueger et al. [54] extract frequent visited places from vehicle movement data and further use semantics distilled from the social network to decode daily activities of people. Approaches putting less focus on visualizations and more on data mining mechanisms have been proposed by some studies [7, 83, 84] to discover human movement patterns based on LBSNs. For the research on collective movement, clustering is a popular approach in looking for common patterns. Andrienko et al. [28] propose a wide range of clustering-based analytics models and combine those with visualization techniques. Their clustering models group similar trajectories as a whole and extract common trips. In this work, we focus on finding com-

mon sub-trajectories. Our clustering of sub-trajectories (as opposed to whole trajectories) enables the extraction of similar portions of trajectories, even when no overall clusters may exist.

Existing anomaly detection models [85–87] for trajectory data have mainly focused on identifying outliers from a target dataset. The models are usually based on non-supervised learning—they generally do not have factors for the outliers, and assume that the outliers make for a small sub-set from the entire dataset. These models look for major patterns and determine whether each trajectory belongs to the majority according to specific criteria. However, during abnormal situations, even the major behaviors can be unusual compared to normal situations.

To address this challenge, our work focuses on the anomalous human behavior analysis through the combination of user expert knowledge and automatic anomaly detection models. The research [33, 44, 88] dealing with GPS data for collective movement analysis takes advantage in high spatial density compared to density of LBSNs. However, it is difficult to collect data for areas of interest and the data usually has no other context. In order to resolve these issues, we utilize additional context (i.e., Tweet text) from LBSNs and visually incorporate the information to enhance the human movement analysis by improving situational awareness.

2.6 Predicting Movement

Recent advances in location acquisition technology have generated fine-grained movement data that contains individual object movement tracking. For predicting movement data, a movement modeling process [89] is first performed where individual moving objects are embedded in a 2D or 3D Euclidean space as a series of locations. Then varying algorithms are applied for predicting future movement of these modeled objects.

There are many approaches proposed for these movement prediction. The most common underlying algorithms utilized are based on a structured prediction model, such as Hidden Markov Models (HMM) [20, 24], Conditional Random Fields [90, 91], and Bayesian

network [92]. For example, Mathew et al. [24] propose a hybrid HMM model where moving object's location histories are clustered, and the model is trained for each cluster with location characteristics as unobservable parameters. Some of previous work uses template matching based on feature extraction and similarity metrics. Another popular approach to predict the next location of a new trajectory utilizes trajectory pattern matching algorithms. Much work can be grouped in this category. Monreale et al. [23] build a trajectory pattern tree and find the best matching pattern to predict the next destination. Ying et al. [21] utilize geographic and semantic context (e.g., bank, park, school) of geo-location points to generate a semantic sequence. Then, they cluster the trajectories based on the semantic sequence—each cluster contains the frequent behaviors of similar users. Finally, they evaluate the next location based on user's semantic behaviors using the pre-defined clusters.

Although many techniques and algorithms have been proposed for frequency-based individual object movement prediction, not much research has been performed on predicting group movement. Recently many visual analytics approach [93–95] is used to support group movement analysis. Andrienko et al. [93] propose a visual analytics framework to support data abstraction and generalization for modeling transportation networks and three high-level tasks, assess, forecast, and develop options. Landesberger et al. [94] contribute another visual approach to reveal temporal changes of human mobility patterns. While they also present the spatiotemporal variation of movements, they do not provide forecasting features. OD-Wheel [95] detect and analyze original-destination dynamic patterns of different regions in the central city. In this work, we embed individual movements into a two-dimensional Euclidean space and model for the space instead of the moving objects. Given a space with a large number of moving objects, we discretize the space into smaller sub-spaces and model the movement flows for each sub-space. Our model forecasts the future directional density of moving objects based on observed historical patterns of each sub-space using a seasonal trend analysis technique [3]. We then combine the results to visualize the future flow as a whole for the entire space. Our model takes into account flow forecasting of a large group of human crowds and also visualization of their continuous flow. Movement analysis as a large group using flow maps can provide a valuable spatial

overview of the group movement [28] while preserving the privacy of individuals [20, 25]. We also provide an interactive visual analytics environment which enables effective exploration of the predicted results and further analysis [89].

3. VISUAL ANALYTICS OF SPATIOTEMPORAL SOCIAL MEDIA FOR ABNORMAL EVENT DETECTION

Social media services, e.g, Twitter, Youtube, Flickr, provide a rich and freely accessible database of user-generated situation reports. As advances in technology have enabled the widespread adoption of GPS enabled mobile communication devices, these reports are able to capture important local events observed by an active and ubiquitous community. The different forms of social media content provided by the users, such as microposts, images or video footage, can have immense value for increasing the situational awareness of ongoing events.

However, as data volumes have increased beyond the capabilities of manual evaluation, there is a need for advanced tools to aid understanding of the extent, severity and consequences of incidents, as well as their time-evolving nature, and to aid in gleaning investigative insights. Due to the large number of individual social media messages it is not straightforward to analyze and extract meaningful information. For example, in Twitter, more than 200 million Tweets are posted each day [96]. Thus, in a developing event, the relevant messages for situational awareness are usually buried by a majority of irrelevant data. Finding and examining these messages without smart aggregation, automated text analysis and advanced filtering strategies is almost impossible and extracting meaningful information is even more challenging.

To address these challenges, we present an interactive spatiotemporal social media analytics approach for abnormal topic detection and event examination [97]. In order to find relevant information within a user defined spatiotemporal frame we utilize the LDA topic model [2], which extracts and probabilistically ranks major topics contained in textual parts of the social media data. The ranks of the categorized topics generally provide a volume-based importance, but this importance does not reflect the abnormality or criticality of the topic. In order to obtain a ranking suitable for situational awareness tasks, we discard daily

chatter by employing the STL [3]. In our work, globally and seasonally trending portions of the data are considered less important, whereas major non-seasonal elements are considered anomalous and, therefore, relevant.

However, due to the large volumes of data, the very specific syntax and semantics of microposts and the complex needs of situational analysis, it would not be feasible to apply these techniques in the form of a fully automated system. Therefore, our whole analysis process, including the application of automated tools, is guided and informed by an analyst using a highly interactive visual analytics environment. It provides tight integration of semi-automated text-analysis and probabilistic event detection tools together with traditional zooming, filtering and exploration following the Information-Seeking Mantra [4].

3.1 Spatiotemporal Social Media Analytics for Event Examination

Since several social media sources recently provide space-time indexed data, traditional techniques for spatiotemporal zooming, filtering and selection can now be applied to explore and examine the data. However, the vast amount of data is beyond the human analytics capabilities. In order to cope with the data volumes, traditional interaction and visualization techniques have to be enhanced with automated tools for language processing and signal analysis, helping an analyst to find, isolate and examine unusual outliers and important message subsets.

To address this issue, we present an interactive analysis process that integrates advanced techniques for automated topic modeling and time series decomposition with a sophisticated analysis environment enabling large scale social media exploration. In part 3.1.1 of this Section we first explain how the Latent Dirichlet Allocation, a well established topic modeling technique in the information retrieval domain, can be used to extract the inherent topic structure from a set of social media messages. The output of this technique is a list of topics each given by a topic proportion and a set of keywords prominent within the topics messages. In a subsequent step, our system then re-ranks the retrieved topic list by identifying unusual and unexpected topics. This is done by employing a seasonal-trend de-

composition algorithm to the historic time series data for each topic, retrieving its seasonal, trending and remainder components. Using a z-score evaluation, we locate peaks and outliers in the remainder component in order to find an indicator of unusual events. While the LDA topic extraction is done primarily for Twitter data, the abnormality estimation is also applied to different social media data sources, such as Flickr and YouTube, for each topic. This is achieved by searching matching entries for each term of a topic and applying the same STL analysis on the resulting time series. The results are available to the analyst for cross validation. The details of this step are described in Subsection 3.1.2 and the complete detection model is formally described in Subsection 3.1.3. In Section 3.2, we describe how powerful tools based on these techniques are used within our analysis environment, Scatterblogs, in order to iteratively find, isolate and examine relevant message sets.

3.1.1 Topic Extraction

Our monitoring component collects space-time indexed Twitter messages using the Twitter-API. The received messages are preprocessed and then stored in our local database.

Rank	Proportion	Topics
1	0.10004	day back school today
2	0.09717	lls bout dat wit
3	0.09443	people make hate wanna
4	0.08226	**earthquake thought house shaking**
5	0.05869	**earthquake felt quake washington**

Table 3.1
An example of extracted topics and their proportions. We extracted topics from Tweets written on August 23, 2011 around Virginia, where an earthquake occurred on this day. One can see that topics consisting of ordinary and unspecific words can have high proportion values, while the earthquake related topics have a relatively low proportion value.

When users of these services witness or participate in unusual situations they often inform their friends, relatives or the public about their observations. If enough users participate, the communication about the situation constitutes a topic that makes up a certain proportion of all messages within the database, or some messages within a predefined area and timespan. In most cases, however, the proportion will be smaller than that of other prevalent topics, such as discussions about movies, music, sports or politics. In order to extract each of the individual topics exhibited within a collection of social media data, we employ Latent Dirichlet Allocation, a probabilistic topic model that can help organize, understand, and summarize vast amounts of information.

The LDA topic model approach, as presented by David Blei et al. [2], is a probabilistic and unsupervised machine learning model to identify latent topics and corresponding document clusters from a large document collection. Basically, it uses a "bag of words"approach and assumes that a document exhibits multiple topics distributed over words with a Dirichlet prior. In other words, the LDA assumes the following generative process for each document: First, choose a distribution over topics, choose a topic from the distribution for each word, and choose a word associated with the chosen topic. Based on this assumption one can now apply a Bayesian inference algorithm to retrieve the topic structure of the message set together with each topic's statistical proportion and a list of keywords prominent within the topic's messages. Table 3.1 shows an example set of extracted topics resulting from the application of LDA to Twitter data ordered by the proportion ranking. The example social media data was collected from Twitter for the Virginia area on August 23rd. On this day, the area was struck by an earthquake with a magnitude of 5.88. As seen in the table, this earthquake event was captured as a topic within the Twitter messages.

In our system, the MALLET toolkit [98] is used for the topic analysis. Prior to the topic extraction, the stemming algorithm KSTEM by Krovetz [99] is applied to every term in the messages. The results of KSTEM are more readable and introduce fewer ambiguities than the often used Porter stemmer.

For the unsupervised LDA classification and topic retrieval one has to define two parameters: the number of expected topics and the number of iterations for the Gibbs sampling

Number of Iteration Steps in the LDA process
50
foursquare pic hall brooklyn time night day back newyork nyc tweetmyjobs finance york brooklyn ave street york ave park btw
300
time back night day york ave brooklyn btw pic bar food nyc **foursquare occupywallstreet park mayor** newyork tweetmyjobs finance citigroup
1000
time night nyc day york ave brooklyn park **foursquare occupywallstreet mayor ousted** newyork tweetmyjobs finance citigroup **san gennaro street italy**

Table 3.2
An example of topic model results depending on the number of iteration steps in the LDA process. The topics are extracted from the Tweets posted in New York City on September 17 and 18, 2011 where the Occupy Wall Street protest movement began and a famous festival, *San Gennaro* occurred. A higher number of sampling iterations provides a better topic retrieval describing the two different events.

process [100], which is used in MALLET for the topic inference. The number of topics that should be chosen depends on the size of the document collection and the required overview level. A small number of topics (e.g., 10) will provide a broad overview of the documents,

whereas a large number (e.g., 100) provides fine-grained results. The number of sampling iterations is a trade-off between computation time and the quality of discovered topics. To illustrate this, Table 3.2 shows the experimental results of the topic model using a varying number of sampling iterations while the number of topics was set to four. The topics were extracted from Tweets posted in New York City on September 17 and 18, 2011, where a large group of protesters occupied Wall Street in New York City. A topic indicating the Occupy Wall Street protests can be seen when using at least 300 iterations. At the time of these protests, there was also a famous annual festival, the *San Gennaro*, occurring in Little Italy. This can only be seen when using at least 1000 iterations. As shown in Table 3.2, the topics with 50 iterations do not indicate any meaningful events. The topics with 300 iterations, on the other hand, consist of more distinguishable classes. Finally, the topics with 1000 iterations obviously point out individual events which happened in the city.

3.1.2 Abnormality Estimation using Seasonal-Trend Decomposition

Abnormal events are those that do not happen frequently and usually they cover only a small fraction of the social media data stream. As shown in Table 3.1, even during an earthquake episode, highly ranked topics consist of ordinary and unspecific words. The third and fourth ranked topics include words indicating the earthquake event of August 2011: *earthquake felt quake washington*. From this observation in the distributions of ordinary and unusual topics over the social media data, it is necessary to differentiate the unusual topics from the large number of rather mundane topics. In order to identify such abnormal topics, we utilize the STL [3]. For each extracted topic of the LDA topic modeling, our algorithm retrieves messages associated with the topic and then generates a time series consisting of daily message counts from their timestamps. The time series can be considered as the sum of three components: a trend component, a seasonal component, and a remainder:

$$Y = T + S + R \tag{3.1}$$

Here Y is the original time series of interest, T is the trend component, S is the seasonal component, and R is the remainder component. STL works as an iterative nonparamet-

ric regression procedure using a series of Loess smoothers [101]. The iterative algorithm progressively refines and improves the estimates of the trend and the seasonal components. The resulting estimates of both components are then used to compute the remainder: $R = Y - T - S$. Under normal conditions, the remainder will be identically distributed Gaussian white noise, while a large value of R indicates substantial variation in the time series. Thus, we can utilize the remainder values to implement control chart methods detecting anomalous outliers within the topic time series. We have chosen to utilize a seven day moving average of the remainder values to calculate the z-scores, $z = (R(d) - mean)/std$, where $R(d)$ is the remainder value of day d, *mean* is the mean remainder value for the last seven days, and *std* is the standard deviation of the remainders, with respect to each topic. If the z-score is higher than 2, events can be considered as abnormal within a 95% confidence interval. The calculated z-scores are thus used as abnormality rating and the retrieved topics will be ranked in the analytics environment according to this estimate.

3.1.3 Detection Model

To conclude this section, we formalize our abnormal event detection model based on the probabilistic topic extraction and time series decomposition.

An abnormal event is associated with a set of social media messages that provides its contents, location, and time-stamp. To detect abnormal events for a given area and timespan, we define a set called *social spacetime* as follows:

$$S = (T, \Delta time, \Delta area, msgs) \tag{3.2}$$

where T is a set of topics, $\Delta time$ is a time period (e.g., one day), $\Delta area$ is a bounded geographical region, and *msgs* is a set of messages. The user selected parameters $\Delta area$ and $\Delta time$ define the analysis context for which all messages are loaded into the analysis system. In this context, the user selects a subset of messages (*msgs*) for which the LDA topic modeling procedure (described in Section 3.1.1) extracts the set of topics, $t_i \in T$. Each topic is defined as:

$$t_i = (M_i, W_i, z_i, Y_i, p_i) \tag{3.3}$$

where W_i is a set of words describing the topic, M_i is a set of relevant messages, z_i is an abnormality score (z-score), Y_i is a time series, and p_i is a statistical proportion of the topic in *msgs*.

For each topic (t_i), our algorithm searches relevant messages (M_i) in the selected area ($\Delta area$) and time period ($\Delta time$) and a predefined time span of historic data preceding $\Delta time$ (e.g. one month). Messages are considered relevant if they contain at least one word in W_i. From M_i a daily message count time series (Y_i) is generated from the timestamps of the messages. The algorithm decomposes Y_i to obtain a remainder component series using the STL and calculates a z-score (z_i) from the remainder series. Lastly, it sorts the topics based on the z-scores.

For cross validation of each topic, we search for relevant entries in Flickr and YouTube by their meta-data that includes titles, descriptions, tags, and timestamps, using the respective APIs. We repeat the steps for generating a time series from the collected timestamps, applying STL to decompose the time series, and calculating the z-score from the remainder component series.

3.2 Interactive Analysis Process

The complete topic extraction, abnormality estimation, and event examination are tightly integrated into a highly interactive visual analysis workbench, that allows an analyst to observe, supervise, and configure the method in each individual step. The following sections introduce the details of this system and describe how the event detection is embedded within a sophisticated analysis process as shown in Figure 3.1.

3.2.1 Social Media Retrieval and Analysis System

Our modular analysis workbench ScatterBlogs was already featured in previous works [61, 75]. It proved itself very useful for fundamental tasks like collection, exploration and examination of individual, as well as aggregated, social media messages. The UI of the system is composed of several interconnected views and the main view houses a zoomable

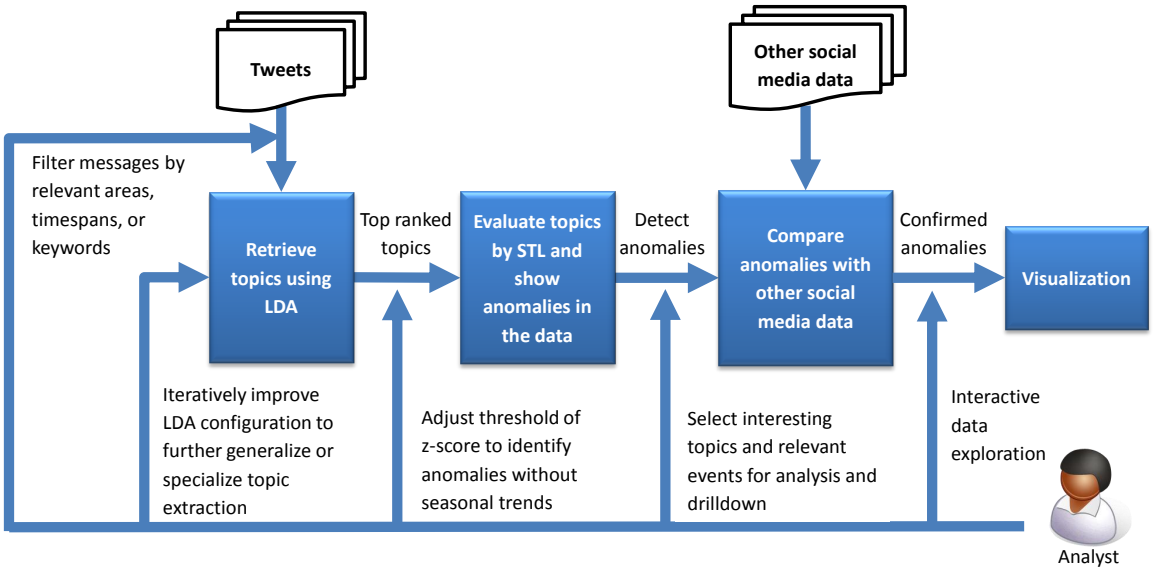

Fig. 3.1. Overview of our iterative analysis scheme for event detection and examination.

openstreetmaps implementation showing message geolocations on a world map. The system features a text search engine and visual content selection tools that can be used to retrieve messages, show spatial and temporal distributions and display textual message contents. Additional visualizations and map overlays provide the analyst with powerful inspection tools, such as a kernel-density heatmap similar to [102], to show aggregated and normalized message distributions and a movable lens-like exploration tool (called 'content lens') that aggregates keyterm frequencies in selected map areas [61]. To indicate spatiotemporal anomalies in the message set, the system features a mechanism to detect spatiotemporal clusters of similar term usage, and suspicious message clusters can be represented as Tag Clouds on the map [75]. For the real-time collection of messages using the Twitter Streaming API the system features a scalable extraction and preprocessing component. This component was used to collect Twitter messages since August 2011 and it currently processes up to 20 Million messages per day, including the almost complete volume of up to 4 million messages that come with precise geolocation information.

3.2.2 Visual Topic Exploration and Event Evaluation

Results from the topic retrieval and event detection as described in Section 3.1 can be iteratively refined by means of visual result presentation and interactive parameter steering. Both, the final result of event detection as well as intermediary findings during data filtering and topic extraction can be used by the analyst to adjust the process in order to identify interesting topics and keyterms as well as relevant map areas and timespans for a given analysis task. New insights can be generated on each of four individual analysis layers which, in conclusion form an iterative analysis loop from data filtering to result visualization:

- **Spatiotemporal Data Filtering:** The analyst selects an initial spatiotemporal context of Twitter messages to be represented in the visualization and to serve as a basis for analysis. He can do so by using textual as well as spatiotemporal query and filter mechanisms that load the relevant base message set from a larger database into active memory. The analyst can further filter the base set and remove unimportant parts by using a time-slider, depicting temporal message densities, or polygon and brush selection tools. Using these tools the analyst can gain an initial impression of the spatial and temporal distribution and location of messages that could be relevant for his analysis task.

- **LDA Topic Examination:** In the subsequent step the analyst can choose to start the topic extraction either on the whole analysis context or on some subset of selected messages. At this stage he can utilize the configuration parameters of LDA extraction to interactively explore available topics by generalization and specialization. In this regard the most important parameter is the number of topics that have to be defined for the topic model inference. If the analyst decreases the number using the provided tools, the extracted topics will be more general. If he increases it, they will be more specific and thus candidates for small but possible important events. Once topics are generated from the data they will be presented to the analyst through a list of small tag clouds for each topic. He can now select the topics from the list to see

their individual message distribution on the map and the temporal distribution in the time-slider.

- **STL Evaluation:** Depending on the analyst's choice, the topics can be evaluated and ordered based either on absolute topic frequency or based on abnormality estimates that have been computed using STL. As described in Section 3.1.2, a valid estimate of abnormality depends on the computation of z-scores from data seven days prior to the observed time frame. Therefore, the STL evaluation will extend the data examination to a range prior to the selected spatiotemporal context, if data is available. Once abnormality is computed for each topic, the topic list will be ordered according to the values and the topics with most outstanding abnormality are highlighted.

- **Crosscheck Validation:** Each selection of messages is accompanied by charts showing the total time series and the remainder components for the selected message set using STL. This is true for spatiotemporal selections as well as for selections using the LDA topic list. In addition to the geolocated Twitter messages this STL is at the same time performed for data that has been extracted from supplemental services like Flickr and YouTube. Based on the multiple charts the analyst can crosscheck the importance and abnormality of examined events and topics.

In our system, the analyst is supposed to iteratively use these means of semi-automated processing, visualization and interaction to refine the selection of messages up to a point where he can begin to examine individual message details. For this task, he can then utilize tools like the content lens for small scale aggregation or the table view to read the messages textual content. The application of these tools is shown in Figure 3.2. Usually the most valuable messages will be reports from local eyewitnesses of an important event or from insiders for a given topic. Thus, to retrieve large quantities of such messages helping to understand an ongoing event or situation will be the final goal of the iterative process. Unusual topics, suspicious keyword distributions and events with high STL abnormality discovered on the repeatedly traversed analysis layers can guide the analysis from a very

broad and general overview to very specific topics and a relatively small message set suitable for detailed examination.

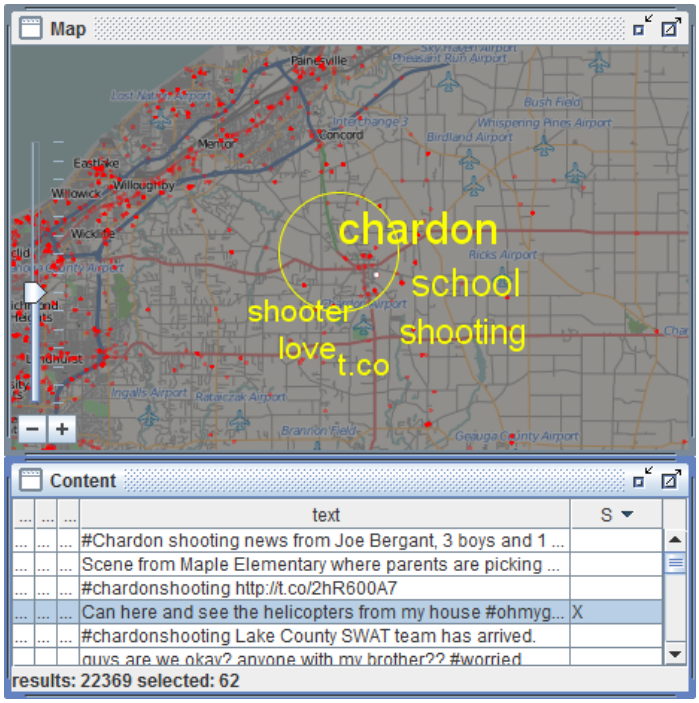

Fig. 3.2. Examining the location of the Chardon high school shooting with a text aggregating content lens.

3.3 Case Study

In this section, we present three case studies for our system covering different types of events including the Chardon High School Shooting, the Occupy Wall Street protests in New York, and the 2011 Virginia Earthquake. The first case shows how analysts can use our system efficiently to find and explore an abnormal event. The second case highlights the differences between social media types by cross validation of a planned event. Finally, the last example showcases the effects of an abrupt, unexpected, natural disaster.

3.3.1 Ohio High School Shooting

On February 27, 2012, a student opened fire inside the Chardon High School cafeteria in the early morning. The gunman killed one student and injured four, from which two eventually died after the incident.

To examine this incident we first locate and select the broader Cleveland area on the map and select a time frame covering three days from February 26 to February 28. Using the text search engine and a wildcard query ('*') we can establish an exploration context showing all messages plotted on the map with their respective contents and meta data listed in a separate table view. First, we want to get a broad overview of the topics discussed in the region and thus we select all messages in the area and apply the LDA extraction tool to the current selection. In order to see the most general topics, we chose a low parameter value for the number of topics and a high iteration count to achieve good separation. At this level of semantic detail, the extracted topics indicate messages about the NBA all-star game (February 26 in Orlando) with keywords like *kobe, game, dunk and lebron* as well as the showing of the movie 'The Lion King'on TV with keywords *king, lion, tv*. If we look at the STL-Diagrams of these topics and the computed z-scores, we also see a peak for these events. By clicking on the retrieved topic representations the associated messages are highlighted in each view. By reading some of the message contents (e.g. *'Watching my fav. Movie on ABC family..... Lion King!!!!', 'Can't wait till the dunk contest starts!'*), the analyst can easily disqualify these from further analysis.

To get a higher semantic resolution we can now increase the number of topics and slightly decrease the iteration count in order to achieve a fast computation. By selecting 20 topics, the topic indicating the shooting event is extracted and indicated by keyterms like *shooting, chardon and school*, alongside the other topics. Although the proportion of the topic is not very high compared to the others, the topic receives a very high z-score (i.e., 3.77) and is ranked among the top five topics (highlighted in orange). Figure 3.3 demonstrates the system view of this observation. An analyst can now select the incident topic to see the spatial distribution of associated messages on the maps as well as the temporal dis-

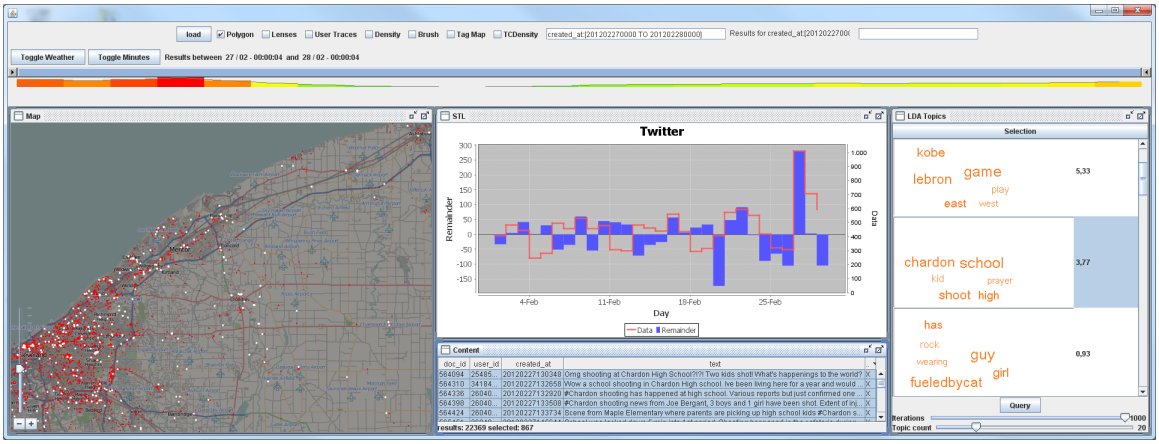

Fig. 3.3. Social media analysis system including message plots on a map, abnormality estimation charts and tables for message content and topic exploration. It can be seen, how the Ohio High School Shooting on February 27, 2012 is examined using the system. The selected messages, marked as white dots on the map, show retrieved Tweets that are related to the event.

tribution in the timeslider histogram. By examining messages using the content lens to aggregate topics over map areas as well as the tools for reading individual message contents, we can easily distinguish between messages informed by media reaction and messages of actual observers in the Chardon High School area. In this case, after isolating the messages from local observers, we find messages like *'Omg shooting at Chardon High School?!?!'* and *'Helicopter overhead. We are on scene. Message from school says students moved to middle school'*.

3.3.2 Occupy Wall Sreet

Starting on September 17, 2011 in the Wall Street financial district in New York City, people have been gathering for the Occupy Wall Street protest movement. The movement against economic inequality has since spread to other major cities throughout the world. Various social media services including Twitter, Facebook, Flickr and Youtube have been utilized both by the participants and the global media for communication and reports about

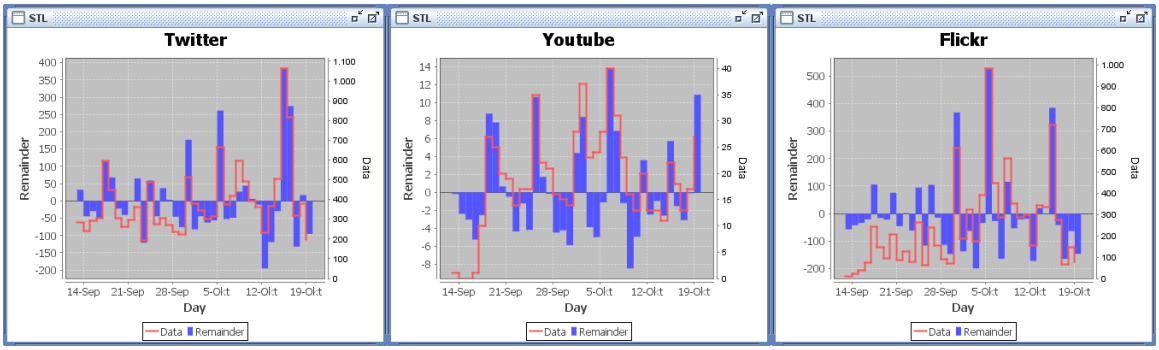

Fig. 3.4. Cross validation of an event using Twitter, Flickr, and YouTube data for the Occupy Wall Street Protests. The protests occurred on Sep. 17 and 30, Oct. 5 and 15. The line charts show the remainder components R (blue) and the original data volumes Y (red) for the STL evaluation. The scales on the right and left side of each chart view are adapted to the maximum values.

the movement in forms of text, images and videos. For the related extracted topic (*occupywallstreet, wall, takewallstreet, takewallst, park*), Figure 3.4 shows the results of our abnormality estimation for the three social media services Twitter, Flickr, and YouTube over the course of one month. As shown in Figure 3.5, in each of the marked regions, at least two of the services show z-scores over 2.0 and they correspond to actual events during the Occupy Wall Street protests. From this experimental result, one can derive a strong correlation between the three social media data sources. The related data volumes and remainder (R) are shown in Figure 3.4 for all three providers.

As shown in Figure 3.5, on September 17 (the first day of the protests with approximately 1,000 participants [103]), only the Twitter stream received an abnormal score while the Flickr and YouTube data artifacts are delayed by 1-3 days. We attribute this initial delay to the simple nature of Twitter usage compared to Flickr and YouTube where the data potentially has to be recorded, edited, and uploaded and is thus more labor intensive. Additionally, eighty protesters where arrested while marching uptown on September 24, but even though Flickr and YouTube reaction on this event created higher z-scores in the following days, they were not significant enough to register an event. The following spikes

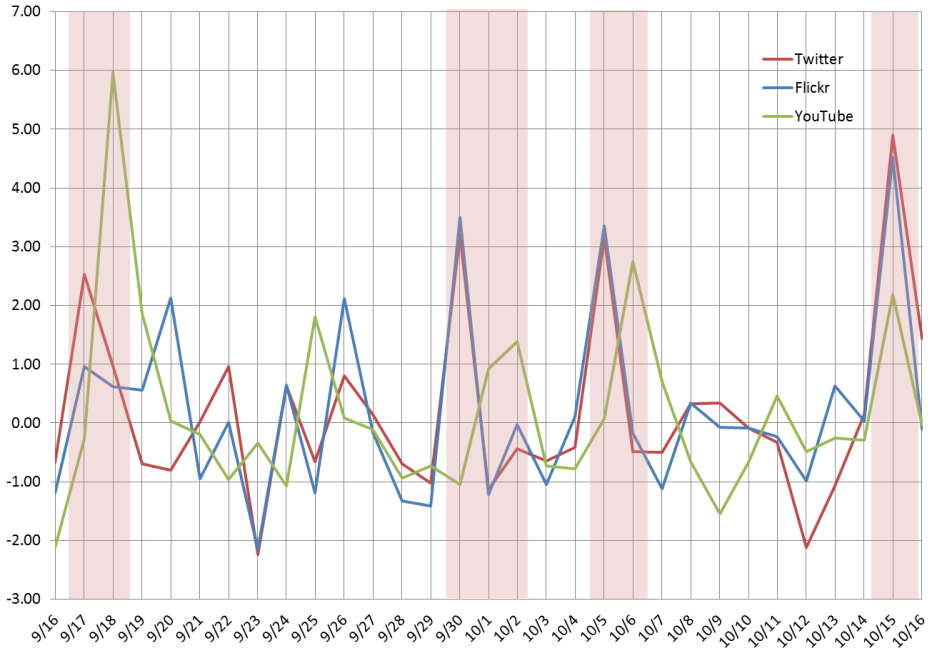

Fig. 3.5. Abnormality and correlation on multiple social media sources. As a result of high z-scores around the same time periods, we found a strong correlation between the three social media sources. Marked regions correspond to periods where at least 2 providers received scores over 2.0.

of high z-scores overlap with a march across the Brooklyn Bridge (Oct. 1 [104]), a large demonstration (Oct. 5 [105]), and globally coordinated protests (Oct. 15 [106]).

3.3.3 2011 Virginia Earthquake

For the last use case we examine a magnitude 5.8 earthquake that occurred on the afternoon of August 23rd 2011 in Mineral, Virginia [107]. Starting with the minute of the earthquakes occurrence, Twitter users posted more than 40.000 earthquake-related Tweets reporting tremors they felt along the East Coast [108]. Among these were messages like: *'EARTHQUAKE!!!!!!!'*; *'Whoa!!!! Just experienced an earthquake here in Virginia!!!!'*; and *'Omg I just felt an earthquake'*. Figure 3.6 gives an impression how our system is applied to examine this event.

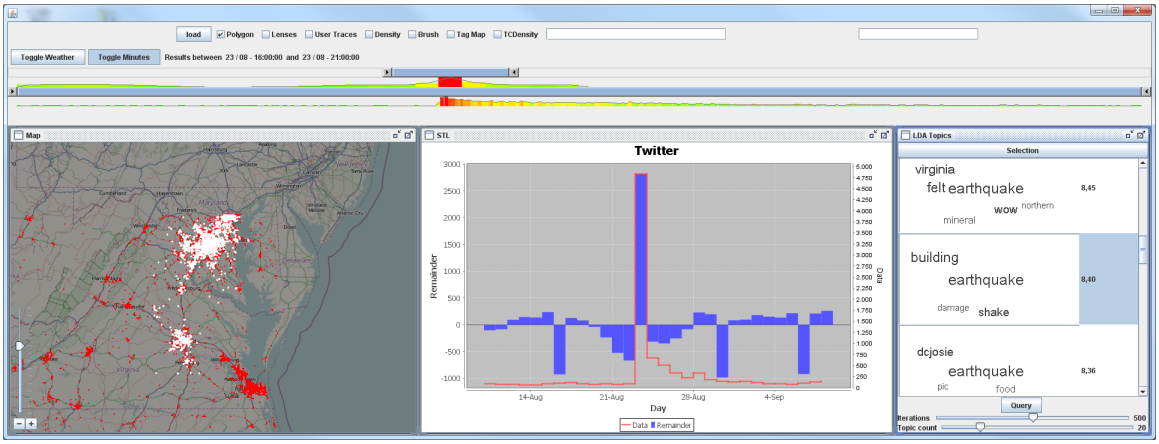

Fig. 3.6. Virginia earthquake on August 23rd, 2011. Our abnormal event detection system detects the earthquake event using our STL based anomaly detection algorithm. The abnormality degree is extremely high on August 23rd, 2011 (times are given in UTC).

For the analysis we begin with selecting the Virginia area from Baltimore to Virginia Beach and three days around the 23th. A topic extraction with 5 topics and just 100 iterations already retrieves two earthquake related topics showing that this event is very prominent within the selection. By clicking these topics one can observe that the highest density of earthquake messages can be found in the Washington, Baltimore and Richmond areas.

To observe the areas in more detail we combine the topic selection with a spatial selection of the three cities and reapply the topic extraction. This time we use 20 topics with 500 iterations. Since we are now operating only on earthquake related messages, the retrieved topics all contain earthquake as a dominant keyword. On this level of detail we can see topics indicating that buildings have been evacuated due to the earthquake (*earthquake, people, evacuated, early, building*) and that damage has been caused (*earthquake, building, shake, damage*). The z-scores for all top ranked topics are now very high (often above 8.0) and thus indicate the high abnormality of this event.

Finally, when going into even higher detail with 100 topics and 1000 iterations we can see smaller events within the big earthquake event. For example, one topic indicates that damage was caused to the Washington Monument and by clicking on the topic we can see messages like *'damage to Washington Monument'*; *'Washington Monument is tilting?!? '*; and *'Helicopter just landed next to Washington Moniment, west side. #DCearthquake '*. There are also misleading messages, indicating that the damage to the Washington Monument was just false rumors: *'the Washington monument was not damaged in any way from the earthquake. #rumor'*. However, media crosschecks show that visible damages did in fact happen and will probably cost the city 15 million dollars to repair [109].

At this point, it is important to note, that while several earthquake topics produced significant z-values in Twitter, the event did not produce high z-scores in Flickr and YouTube. This is probably due to fact, that many people will write a quick message after a shock has been felt by themselves, but it takes quite some time until images or videos are uploaded from cameras to Flickr and YouTube. The event also demonstrates that large and unexpected events will produce immediate and significant reactions in services like Twitter and they can thus easily be detected by using our system.

3.4 Discussion

In this section we want to discuss four important notes and observations relevant to the presented approach.

Event Types: As was demonstrated with the three case studies, events in social media can be categorized into two different types. The 2011 Virginia Earthquake and the Ohio High School Shooting can be categorized as abrupt or disaster events, while Occupy Wall Street can be considered a social and planned event. The two types of events have quite distinguishable features. For the abrupt events, there is a strong change in daily counts mainly in the text based Twitter messages. For the planned event, the Twitter signal may still be faster, but due to the gradual increase and decrease, it is less pronounced. In contrast, Flickr and YouTube have delayed, but very prominent changes, for planned events;

however, we could not find significant signals for abrupt events. This reflects that video and photo recording happen rarely during abrupt events. Social events, e.g., Occupy Wall Street or election debates, however, have a high impact on such multimedia based social media; Relevant videos, photos, and even meta-data (e.g., descriptions, tags) allow analysts to find additional information about them. We, therefore, think that cross validating events among multiple social media types is important in order to establish situational awareness.

Base Data: Regarding the base data, it is important to note, that our approach depends on geo-located Twitter messages with precise coordinates, which are only a fraction of the whole Twitter stream. While this fraction still consists of several million messages per day, it is not a representative sample of the population, because it mainly covers mobile users equipped with GPS enabled devices. We think, however, that mobile users, who share their daily experiences freely, are the most relevant group for situational awareness scenarios. Some studies [110, 111] tried to overcome the problem of location information scarcity in Twitter messages, which adds another source of uncertainty. First, the user's self reported locations can be outdated. Second, the geo-coding of the location can be considerably wrong due to place name ambiguities. Furthermore, we have just shown the feasibility of the approach for Twitter, Flickr, and YouTube data, but it can easily be adapted to other social media providers like Facebook or Forsquare as well, in order to widen the sample of the population.

Probabilistic Models: In this work, we use STL to decompose time series of topic streams. There are many alternative statistical models for this task, such as DHR (Dynamic Harmonics Regression) [112] and SARIMA (Seasonal AutoRegressive Intergrated Moving Average) [113]. DHR and SARIMA models are particularly useful for forecasting and STL can also be used for prediction based on seasonal (periodic) time series [114]. Our main reasons for choosing STL was the fact that it is non-parametric, can be computed faster than SARIMA [114] and needs less training data for equally good results.

End User Feedback: We requested informal feedback from users within our institutes and received comments and suggestions. To compare the LDA topic modeling plus the seasonal-decomposition based abnormality analysis versus only the LDA topic modeling,

we enabled our system to switch between these modes. The users were impressed by the fact that both results (two lists of topics) from two different modes were quite different. Highly ranked topics by LDA topic modeling consisted of ordinary words, while the combined analysis was indicating unusual events. They noted that the tightly integrated visual analysis workbench was useful to apply the automated methods. Furthermore, they suggested a function allowing people to see a pattern of abnormality for a user-defined topic.

3.5 Summary

We presented an interactive abnormal event detection and examination system for the analysis of multiple social media data sources. The system uses an abnormality estimation scheme based on probabilistic topic modeling and seasonal-trend decomposition to find and examine relevant message subsets. This scheme is tightly integrated into an highly interactive visual analytics system, which supplements tools based on automated message evaluation with sophisticated means for parameter steering, filtering and aggregated result set exploration. Three use cases demonstrated the visualization and user interaction within the system and its capabilities to detect and examine several different event types from social media data. The ability to crosscheck findings based on three distinct social media sources revealed the kinds of correlations that can be expected from various event types.

4. VISUAL ANALYTICS OF MICROBLOG DATA FOR PUBLIC BEHAVIOR ANALYSIS IN DISASTER EVENTS

In this chapter, we introduce a visual analytics approach for public behavior analysis in disaster events. Analysis of public behavior plays an important role in crisis management, disaster response, and evacuation planning. Unfortunately, collecting relevant data can be costly and finding meaningful information for analysis is challenging. A growing number of LBSN services provides time-stamped, geo-located data that opens new opportunities and solutions to a wide range of challenges. Such spatiotemporal data has substantial potential to increase situational awareness of local events and improve both planning and investigation. However, the large volume of unstructured social media data hinders exploration and examination. To analyze such social media data, our system provides the analysts with an interactive visual spatiotemporal analysis and spatial decision support environment that assists in evacuation planning and disaster management. We demonstrate how to improve investigation by analyzing the extracted public behavior responses from social media before, during and after natural disasters, such as hurricanes and tornadoes.

This study is performed using Tweets, as Twitter has been the most popular microblog service in the United States. We extend our previous work [115] with additional features of our system and examine their capabilities with several expanded examples in Section 4.2.2. We also add a discussion section for comparisons and analysis of the case studies.

Our system evaluates visual analytics of spatiotemporal distribution of Tweets to identify public behavior patterns during natural disasters. The main features of our approach are as follows:

- **Spatial analysis and decision support:** The system provides effective analysis for exploring and examining the spatial distribution of Twitter users and supporting spa-

tial decision-making using a large volume of geo-located Tweets and multiple types of supplementary information during specific time periods (i.e., disaster events).

- **Temporal pattern analysis:** Our visualization system enables the analysts to analyze the temporal distribution of the number of Twitter users posting Tweets in a given location and time.

- **Spatiotemporal visualization:** We provide a visualization that allows the analysts to simultaneously analyze both aspects: space and time in a single view.

4.1 Problem Statement and Interactive Analysis Process Design

Analysis of public behavior, such as how people prepare and respond to disasters, plays an important role in crisis management, disaster response, and evacuation planning. Recently, social media becomes popular and people utilize it for communications not only in their daily lives, but also in abnormal disastrous situations. Thus, Location-based Social Networks services offer a new opportunity for enhancing situational awareness during disaster events. Unfortunately, collecting relevant data can be costly and finding meaningful information from the huge volume of social media data is very challenging. Therefore, there is a need for an advanced tool to analyze such massive ("big") streaming data and aid in examining the analysis results to better understand situations more efficiently.

Our proposed visual analytics approach provides multiple analysis methods: spatial analysis, spatial decision support, temporal pattern analysis, abnormal topic analysis, and interactive spatiotemporal visualization as shown in Figure 4.1. In our system, all methods are tightly integrated based on a user-centered design in order to enhance the ability to analyze huge social media data (Figure 4.1 (A, B, C)). Our Tweet collection component obtains real-time Tweets using the Twitter API—to collect about 2.2 million geotagged Tweets within the United States per day. In general for spatial analysis, the required accuracy of the geocoordinate depends upon the required level of location granularity. The data, however, is generated by very reliable GPS and software. We can be reasonably certain about the data accuracy as illustrated in [116]. For the temporal accuracy of Tweets,

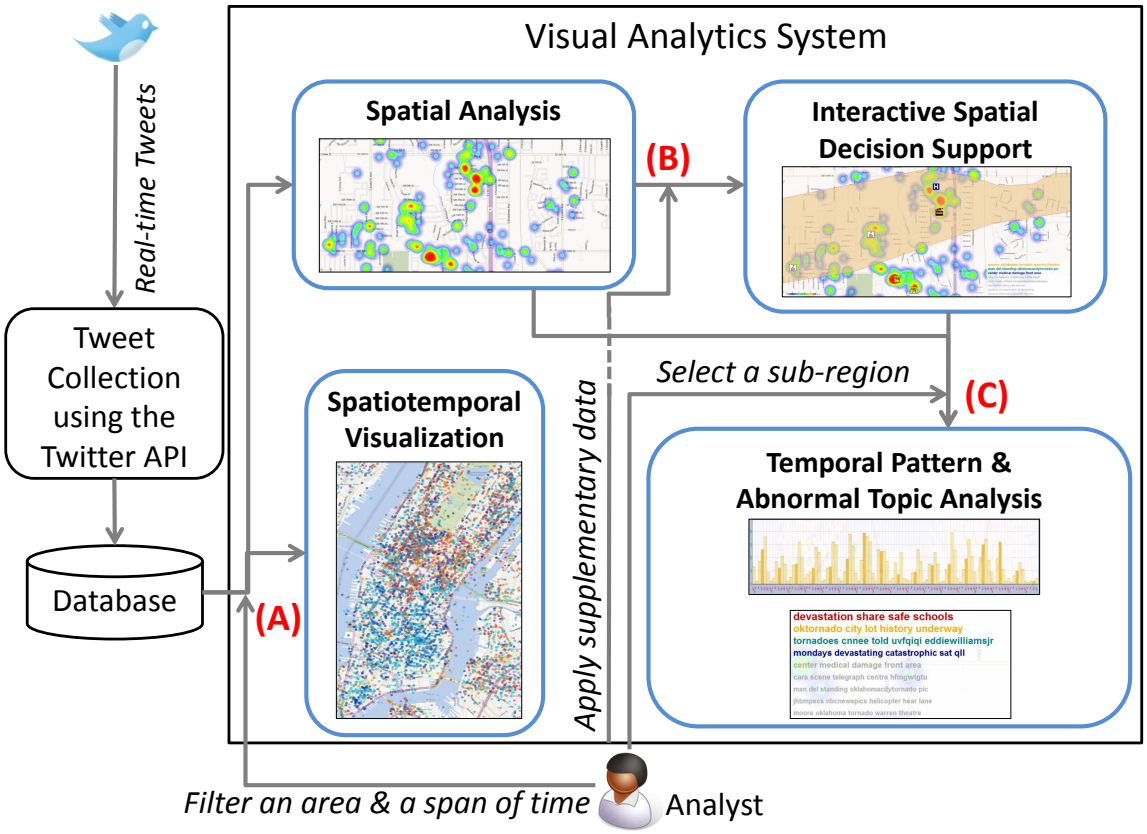

Fig. 4.1. Overview of our interactive analysis scheme for public behavior analysis using social media data.

we use the time when each Tweet is created. Therefore, it is highly accurate if the time setting of the device posting a Tweet is correct. This large volume of data is stored in our database in order to maintain and track the history of the Twitter stream. Our system allows the analysts to query Tweets with a specific area and time span condition (Figure 4.1 (A)). The initially selected spatiotemporal context of Tweets can be represented by two different analytics components: spatial analysis and spatiotemporal visualization. Spatial analysis allows the analysts to examine the overall distribution of Twitter users and discover hotspots where relatively more Twitter users post Tweets. The analysts are able to add supplementary information (infrastructure locations, tornado paths) on top of current information representing outcomes in order to better understand events and increase situa-

tional awareness (Figure 4.1 (B)). Furthermore, the analysts can select a sub-region within the initial area, so that he can analyze the temporal patterns of the number of Twitter users and extract abnormal topics from the text messages in the selected region (Figure 4.1 (C)). In addition, our interactive spatiotemporal visual analytics provides a single view representation for the analysis of both aspects: spatial and temporal characteristics of Tweets at the same time.

4.2 Spatiotemporal Analysis

In this work, we present a visual analytics approach to handle the vast amount of microblog data such as Twitter messages, provide interactive spatiotemporal analysis, and enable the use of multiple types of supplementary spatial infrastructure information for spatial decision support. Analysts select an initial spatiotemporal context of Tweets to be represented in the visualization to serve as a basis for analysis. They can also perform the interactive spatiotemporal queries that load the relevant datasets from a larger database.

4.2.1 Spatial Analysis

Social media embedding geo-location information into the data is extremely useful in analyzing location-based public behaviors. Such spatial analysis, therefore, is important in order to manage and prepare plans for disaster and emergency situations.

In late October in 2012, a massive hurricane, Sandy, devastated Northeastern United States [117]. Due to the severeness of the hurricane, on October 28th in 2012, the New York City Authorities ordered residents to leave some low-lying areas—the mandatory evacuation zones (red color) are shown in Figure 4.9 (Right). We investigate an area of Manhattan, since the area is the most populated and severely damaged. Through the map view in our system, analysts navigate to the Manhattan area in New York City and filter Tweets posted within the area. Initially we tried to reveal public movement flows during the disaster event, but the movement patterns were too complicated to find meaningful flows due to movement randomness and the visual clutter of the flows. Then, we examined

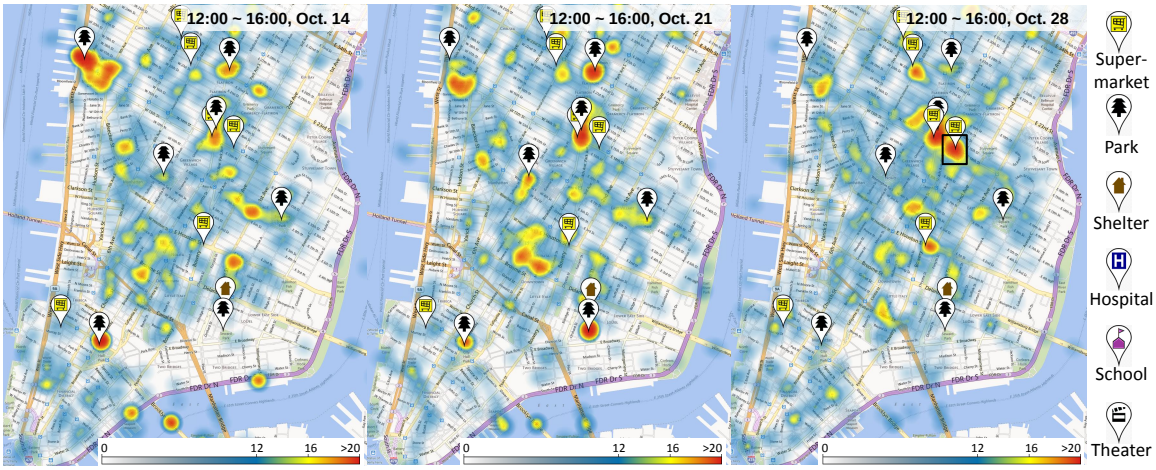

Fig. 4.2. Spatial user-based Tweet distribution in the Manhattan area in New York City during four hours right after the evacuation order (from 12:00 PM to 4:00 PM on October 28th, 2012 (Right)). Previous distribution of Tweets on 14th (Left) and 21st (Center).

the spatial distribution of the users for specific time frames. Based on our experiments, a geospatial heatmap was useful for an overview of the spatial distribution and for trend approximation. We utilize a divergent color scheme to generate the heatmap, where saturated colors are used for the data distribution to avoid any confusion from the color scheme from the desaturated colormap of the background map. Analysts can specify a threshold range to emphasize hotspots, where the upper bound is mapped to a red color and the lower bound to a yellow color. Additionally, the blue color is mapped by the analysts to the value of the overall distribution of Twitter users. In Figure 4.2, we show three heatmaps of spatial user-based Tweet distribution from 12:00 PM to 4:00 PM on October 14th (Left), 21st (Center), and 28th (Right). In this work, we use the number of Twitter users instead of the number of Tweets for the heatmap generations to properly reflect the flow of evacuation unbiased by personal Tweet activity or behavior of individual users, since some enthusiastic Twitter users generate a large number of Tweets at the same location during a short time period (more than 20 Tweets per hour). The heatmaps in Figure 4.2 (Left and Center) represent normal situations of Twitter user distribution in the Manhattan area, and the

heatmap (Right) shows the situation right after the evacuation order that was announced at 10:30 AM on October 28th, 2012. This standard heatmap visualization allows analysts to explore the spatial pattern of Twitter users for any specified time period. In Section 4.2.2, we will provide further analysis for the spatial decision support.

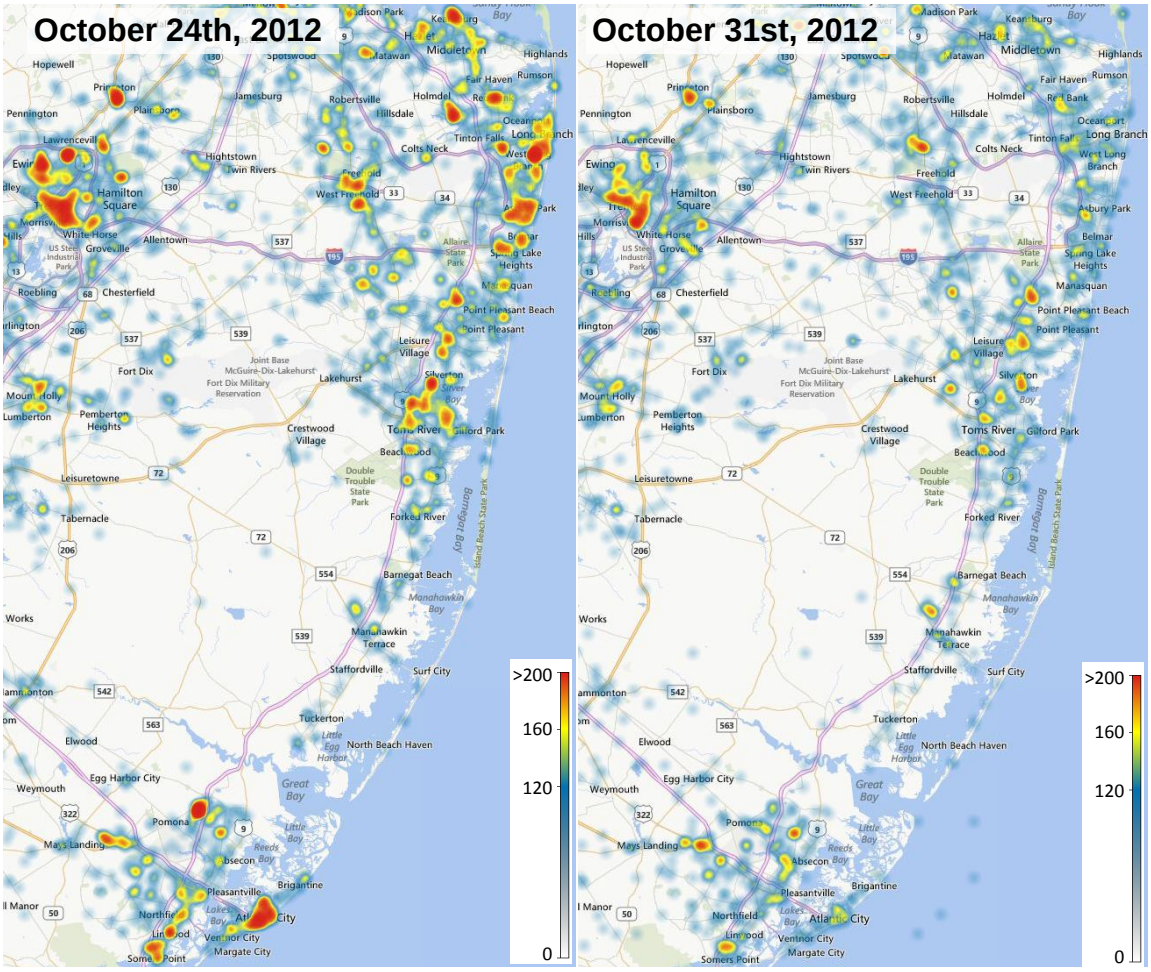

Fig. 4.3. Twitter user distribution on the eastern coast area in New Jersey, after the hurricane passed over the area on October 31st (Right). Previous distribution on October 24th is shown on the Left.

Hurricane Sandy damaged not only New York City, but also the entire eastern coast area of New Jersey. Most cities in the area also announced evacuation orders on October 28th, 2012. The distribution of Twitter users in the area from Atlantic City to the upper eastern

shore area for two different dates are shown in Figure 4.3. The heatmaps in Figure 4.3 (Left) represent the previous normal situation of Twitter user distribution on October 24th and the heatmap (Right) shows the post distribution after Sandy passed over the area on October 31st. As shown in the result, many hotspots are gone or diminished. This situation shows that the number of Twitter users had significantly decreased after the hurricane damaged the area. In fact, a huge number of homes were damaged or destroyed and a couple of million households lost power because of Hurricane Sandy [118]. In disaster management this type of visualization can support analysts estimating which areas were highly damaged and even which areas still need reconstruction.

4.2.2 Spatial Decision Support

In Section 4.2.1, we introduced our spatial analysis to explore the Twitter user distribution. In addition to the analysis, our system allows the analysts to utilize supplementary information in order to support understanding of the situations and decision-making in disaster management. The spatial characteristics together with heterogeneous information can assist in disaster management and migrating hazards where the problems have spatial components [119]. The supplementary information can be various types of infrastructures (i.e., school, park, supermarket, and shelter), as well as spatial information of disaster events (i.e., hurricane path and damage area of a tornado). In this section, we describe how our system supports spatial decision-making by correlating such spatial information with location-based microblog data.

Infrastructure Data

During a natural disaster event, such as Hurricane Sandy, analysts would assume that many people might want to go to the supermarket before staying or evacuating, but they would need supporting evidence before making appropriate decisions and plans. With our system support, the analysts can simply overlay the locations of large supermarkets on the heatmap of the Twitter user distribution. The infrastructure locations are indicated

by standard symbols [120] as shown on the right side of Figure 4.2. A relatively large number of people immediately went to supermarkets nearby the evacuation area, instead of the emergency shelter as shown in Figure 4.2 (Right). However, October 28th was Sunday and many people generally would go for grocery shopping on Saturday or Sunday; therefore, the analysts might need to verify whether the heatmap shown in the figure is a normal periodic situation. The analysts can investigate new Twitter user distributions for different time frames by simply manipulating the time context. In Figure 4.2 (Left and Center), we show two distributions for one and two weeks before the disaster period respectively. Here, we see that the hotspot locations are very different from the ones for October 28th shown in Figure 4.2 (Right). For further analysis, we can explore another popular Sunday location—large parks—by superimposing the locations on each heatmap. As shown in Figure 4.2 (Left and Center), many hotspots overlap with the park areas in normal situations. Therefore, we can conclude that the situation on October 28th is an unusual non-periodic pattern.

Disaster Event Data

In Section 4.2.2, we explained how the infrastructure data help the analysts to understand and examine the emergent situations. During severe weather conditions, people tend to be sensitive to the dynamic variance of the weather conditions. Relationship analysis, therefore, between the public responses and the spatiotemporal pattern of the severe weather is important. Our system overlays geographic information of disaster events, for example, center positions and tracks of a hurricane, and damaged areas by a tornado, in order to provide further analysis. Two case studies are presented as follows:

Track of Hurricane: Figure 4.4 (1) and (2) show the southeastern coast areas of the United States, whereas, Figure 4.4 (3), (4), and (5) show the northeastern coast areas. In the figures the distributions of Twitter users for each consecutive date, from October 26th to 30th, 2012, are presented using the heatmap visualizations. We use the number of Twitter users who posted Twitter messages containing one of the following keywords:

hurricane, storm, and *sandy* in order to analyze Tweets that are highly related to Hurricane Sandy. Note that Hurricane Sandy reached the southeastern Florida coast on October 26th and passed, then, over the northeastern coast on October 30th, 2012 [117]. As shown in Figure 4.4, our system is able to overlay the track of the hurricane on the map. The blue pins and the blue lines represent the center locations of the hurricane and its path respectively.

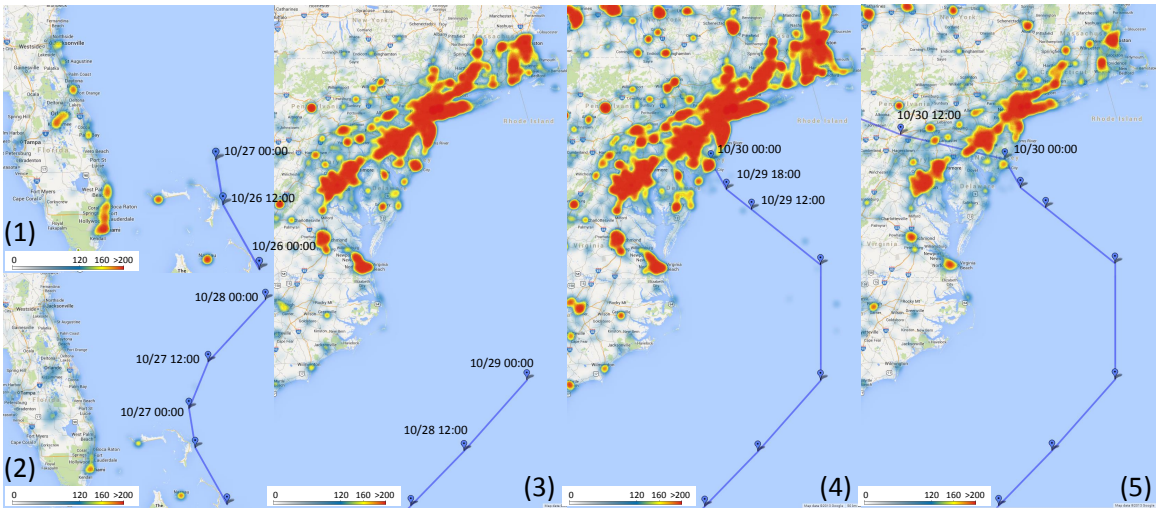

Fig. 4.4. Distribution of Twitter users of each consecutive date (Oct. 26 $\sim$ 30, 2012), who post hurricane related Tweets on the southeastern (1 and 2) and northeastern coast (3, 4, and 5) area of the United States. We can see the variance of Twitter user reactions along the track of the hurricane center locations.

Twitter users also actively respond to the severe weather conditions. In Figure 4.4, we indicate that the distribution pattern of Twitter users had dynamically varied along the track of the hurricane center locations. When Sandy moved to the southeastern coast on October 26th, there were bursts on eastern Florida's coast (Figure 4.4 (1)). Next day, the bursts disappeared, because Sandy moved towards the northeast away from the east coast of United States (Figure 4.4 (2)). Sandy kept moving towards a few hundred miles southeast of North Carolina on October 28th (Figure 4.4 (3)). In the next day, the hurricane's track bent towards the north and the hurricane made landfall at night in the northeast of Atlantic City (Figure 4.4 (4)). Throughout the days, Twitter users were actively reacting to Hurricane

Sandy' arrival in a wide range of areas. After the landfall, the storm turned toward the northwest and was gradually weakened. The big outbreaks were diminished on October 30th as shown in Figure 4.4 (5). As shown in the figures, we can see how Twitter users reacted according to the spatiotemporal pattern of the severe weather conditions in the social media domain.

Damage Area from a Tornado: An extremely strong Tornado passed through the city of Moore in southern metropolitan Oklahoma City [121] in the afternoon on May 20th, 2013. The larger than one-mile-wide tornado damaged the city with a wind speed of more than 200 mph. Figure 4.5 shows the damaged part of the city. The tornado entered the area at about 3:16 PM and exited the area after about 10 minutes. We visualize the distribution of Twitter users on the map during 24 hours, from May 20th 4:00 PM to 21st 4:00 PM. We also overlay an approximate extent of tornado damage (transparent orange color) and locations of multiple infrastructures, such as schools, hospitals, and supermarkets, on the map view. Since the tornado suddenly happened and disappeared, we were not able to find significantly abnormal patterns before and during the event. After the disaster event, however, many Twitter users moved toward some specific areas: two elementary schools, a medical center, a theater, and two large supermarkets. The two elementary schools, the medical center, and the theater were located within the highly damaged area and they were severely destroyed. Also many people were hurt and died in these infrastructures. The increased number of Twitter users was probably due to the fact that many people went to these places in order to rescue the victims [122]. Moreover, people might have gone to supermarkets to obtain indispensable things. In Figure 4.5 (1), the heatmap shows a normal situation of Twitter user distribution in the same area. The distribution is very different from the situation after the tornado hit the area. This example demonstrates how our visual analytics system enables the analysts to analyze public responses using spatial disaster data and infrastructure data for disaster management.

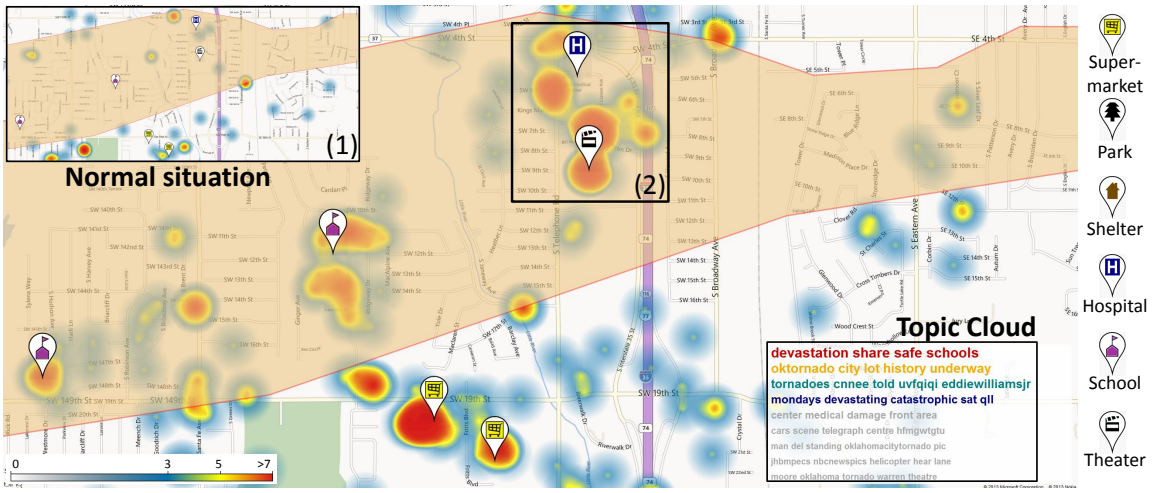

Fig. 4.5. Spatial pattern of Twitter users during 24 hours in the city of Moore after damages from a strong tornado. Relatively many people moved to severely damaged areas after the disaster. This situation is much different from the previous normal situation (1). We selected a specific region (2) that includes severely damaged areas in order to extract topics (3) from Tweets within the selected area.

Fig. 4.6. Topic cloud: Topics from Tweets within the selected area in Figure 4.5 (2) are ordered by their abnormality scores.

Abnormal Topic Analysis

Our system also provides analysts with abnormal topic examination within the microblog data. Each Twitter message provides not only spatiotemporal properties, but also

textual contents. The text messages are also important to understand and examine the emergent situations. Our system allows the analysts to extract major topics from many Tweets posted within a specific area using the LDA [2]. We also employ, then, the STL [3] to identify unusual topics within the selected area. For each extracted topic of the LDA topic modeling, our algorithm retrieves messages associated with the topic and then generates a time series consisting of daily message counts from their timestamps. The time series can be considered as the sum of three components: a trend component, a seasonal component, and a remainder. Under normal conditions, the remainder will be identically distributed Gaussian white noise, while a large value of the remainder indicates substantial variation in the time series. Thus, we can utilize the remainder values to implement control chart methods detecting anomalous outliers within the topic time series. We have chosen to utilize a seven day moving average of the remainder values to calculate the z-scores. Note that we use the z-score as the abnormality score in this work. If the z-score is higher than 2, events can be considered as abnormal within a 95% confidence interval. The details of these techniques are described in the previous work [97]. We select a sub area in Figure 4.5 (2) that includes severely damaged areas: the selected region (black rectangle) on the map. The extracted topics, which are ordered based on their abnormalities, are displayed as Topic Clouds at the bottom-right corner (Figure 4.5 (3)) on the map. The topic cloud is enlarged and shown in Figure 4.6. In this case study, most topics are related to the disaster event. However, the last topic—*moore, oklahoma, tornado, warren, theatre*, has a relatively low abnormality although they seem related to the disaster event, because tornadoes frequently occur in the area. Figure 4.7 shows an abnormality graph for the first topic in Figure 4.6. The abnormality score for the topic had significantly increased when the tornado hit the region on May 20th (Marked region). As shown in Figure 4.7, the abnormality score (6.75) is much higher than the average abnormality score(0.42); therefore, the analysis of the microblog data provides a statistically significant difference during this severe weather condition.

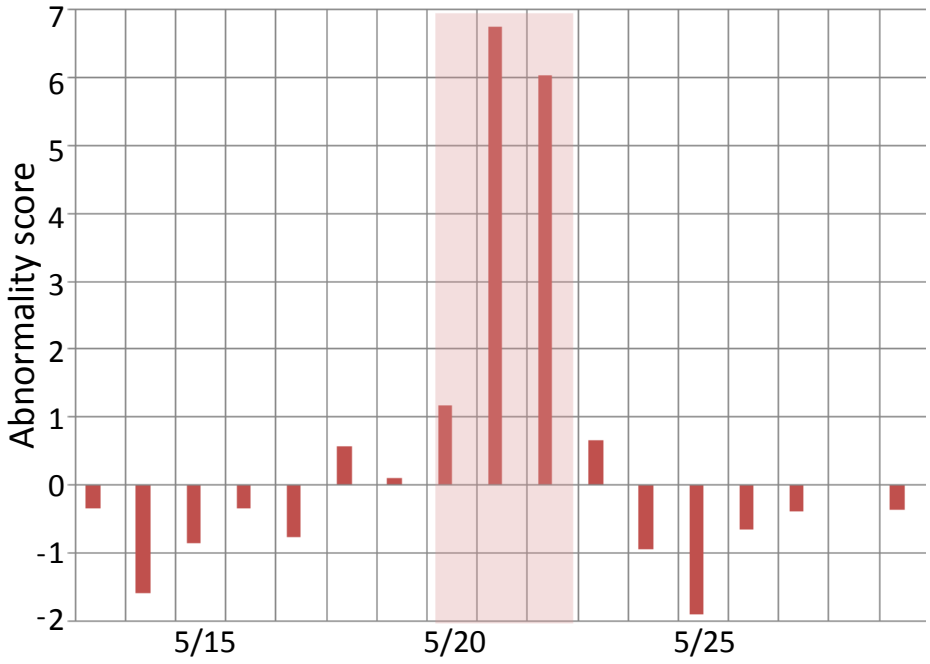

Fig. 4.7. Abnormality of the first topic in Figure 4.6. The abnormality score of the topic had significantly increased when the tornado hit the region on May 20th (Marked region).

4.2.3 Temporal Pattern Analysis

In the previous sections, we presented the spatial analysis of social media and spatial decision support. In this section, we demonstrate analysis of the relationships between the temporal patterns of the number of Twitter users and certain public situational behaviors: how many people go where and how different is it from previous situations? Analysis of temporal trends and relationships between data values across space and time provides underlying insights and improves situational awareness [123, 124].

After selecting the initial spatiotemporal context of Tweets as a basis for the analysis, the analysts can explore the temporal patterns of the number of Twitter users who posted Tweets within the spatial boundary using the bar chart as shown in Figure 4.8. The values of each bar are the number of users in four hour intervals and represent data two weeks before and after the selected date. Once a mouse cursor hovers over one of the bars in the graph, every bar that corresponds to that time period, is highlighted in dark yellow

color as shown in Figure 4.8. As previously mentioned, the heatmap in the figure shows the Twitter user density distribution from 12:00 PM to 4:00 PM on October 28th, right after the announcement of the evacuation order. We select a hotspot that includes one of the supermarket locations: the selected region (black rectangle) on the map in Figure 4.2 (Right). We can indicate that the number of Twitter users (red rectangle in Figure 4.8) in the corresponding time period is higher than for the same time period from other dates (October 14th, 21st and November 4th, 5th) by 35% more from the average. Moreover, there is another interesting finding—the number of people during each of the following time frame (4:00 $\sim$ 8:00 PM) on the dates from the previous weeks are higher than the number of people in the selected time frame. This is because many shoppers were lining up at stores and emptied the shelves to prepare for Hurricane Sandy. Some actual Twitter messages posted in the area are following: *'The line at Trader Joes is unbelievable ...'* and *'There is amazing line here ...'*. Furthermore, since October 29th, the number of people has significantly decreased because most residents left the area before the arrival of the hurricane. The increase in the number of people after one week reflects that some people came back to the area.

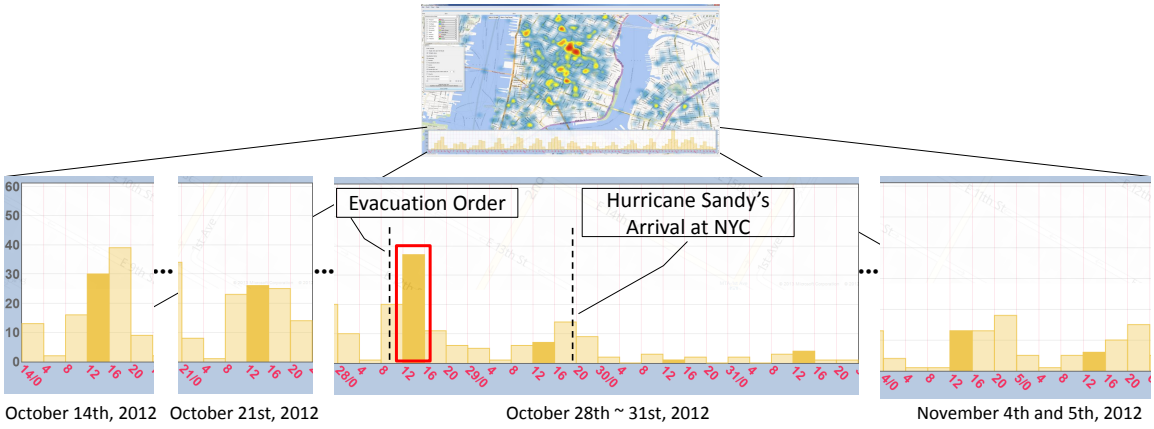

Fig. 4.8. Temporal analysis for public behaviors during the disaster event, Sandy. Top shows our entire system view. The bar chart (Bottom) for the number of Twitter users within the selected region including a supermarket in Figure 4.2 (Right) in four hour intervals is shown. We see that many people went to the supermarket right after the evacuation order.

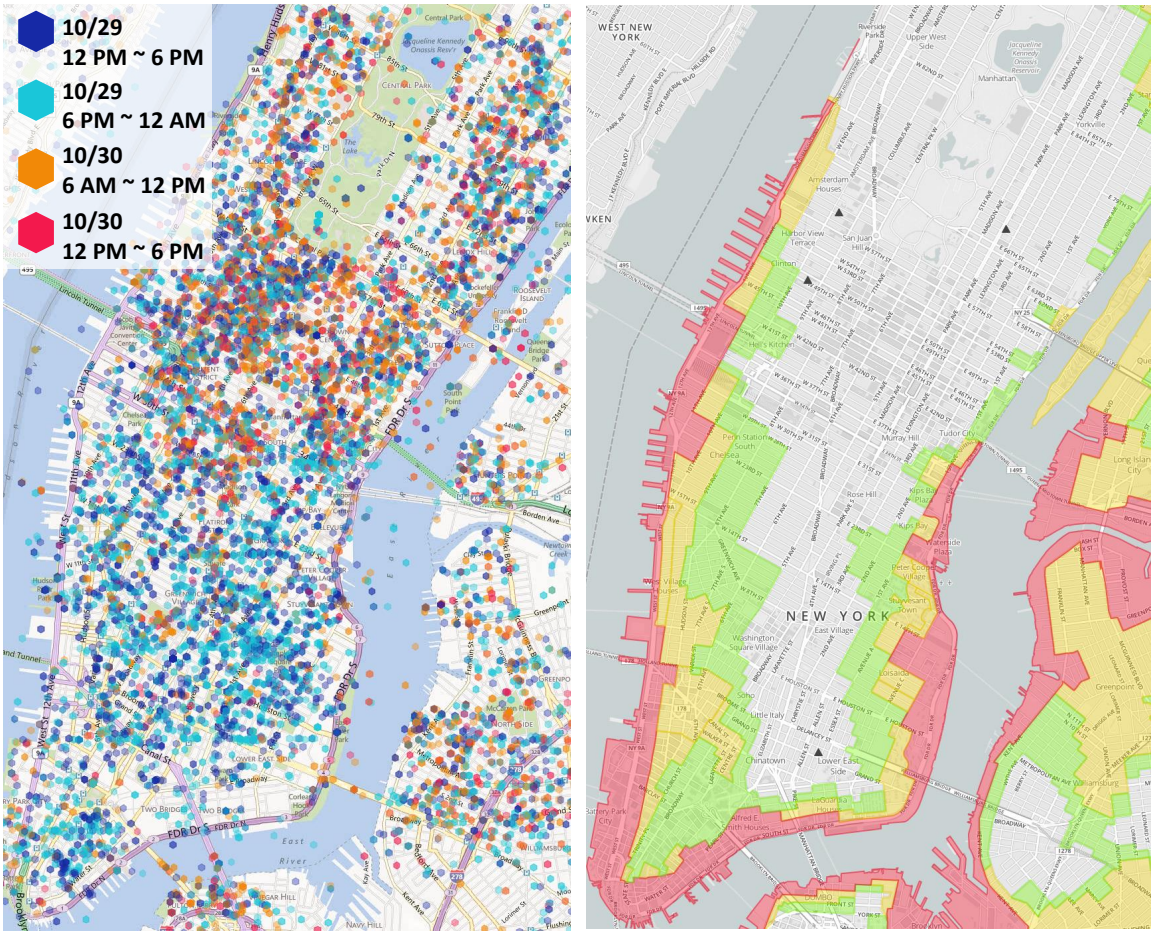

Fig. 4.9. Visualization for spatiotemporal social media data (Left). A hexagon represents the spatial (position) and temporal (color) information of a Tweet. Hurricane evacuation map [125] (Right). Residents in Zone A (red) faced the highest risk of flooding, Zone B (yellow) and Zone C (green) are moderate and low respectively.

4.2.4 Spatiotemporal Visualization

There is abundant research published on the topic of spatiotemporal data visualization. Still, exploration of time-referenced geographic data is still a challenging issue [126]. We introduce a modest visualization that enables analysts to analyze both aspects: space and time in a single view. Each Tweet is independent and contains multiple properties, such as location, time, the number of re-Tweet, etc. In this study, therefore, we utilize a glyph-

based visualization to depict both location and time aspects of the independent data record using two visual features. As shown in Figure 4.9 (Left), each hexagon corresponding to a Tweet represents the spatial and temporal information where the center of each hexagon is the location of each Tweet and the color represents its posting time. In other words, space and time properties are encoded in a single visualization to harness the spatial analysis features of human visual perception [127]. In Figure 4.9 (Left), the hexagons with blue (12 PM $\sim$ 6 PM) or green (6 PM $\sim$ 12 AM)) color correspond to Tweets published on October 29th, 2012 and ones with orange or red color correspond to Tweets posted on the following day after the hurricane. New York City announced the evacuation of Zone A (red color) in Figure 4.9 (Right); residents in Zone A faced the highest risk of flooding, whereas, Zone B (yellow color) and Zone C (green color) are moderate and low respectively. In the visual representation, analysts can indicate overall spatiotemporal patterns of people and their movements during the disaster event—many people still remained at home one day after the mandatory evacuation order, but most people left home on the following day as the hurricane damaged the city.

4.3 Discussion and Evaluation

In this work we found out that the public responses to disaster events in social media streams are different according to the disaster event types. Hurricane Sandy had a long time duration—more than one week, and affected a wide range of areas. Therefore, there were many reactions in the potential damage area before the hurricane impacted the area. However, no or significantly less hotspots were found right after the hurricane passed over the area. This was because the hurricane severely affected the areas—communication facility damage and power outages occurred in the area. Moreover, we found out that unusual post-event situations in the Twitter user distribution continued for a certain time period from a couple of days to more than one week as shown in Figure 4.4 and 4.8. The analysts could estimate how long it took for the reconstructions in the areas.

Regarding the tornado case, we intended to find abnormal patterns in the Twitter user distribution before and during the disaster event but there was no unusual patterns in the area. In contrast to the hurricane, the tornado generally affected the areas relatively shortly, for example, a few minutes to an hour. The abrupt natural disaster did not strongly influence the social media stream before and even during the event. However, as shown in Figure 4.5, we were able to find many hotspots within the damaged areas after the tornado passed. In fact, the tornado damaged some small areas (i.e., a couple of miles wide), in contrast to the wide range of damaged areas for the hurricane case. This indicated that communication facilities were still available and many people were interested in the disaster, similar to the hurricane. Thus, our social media analysis could support the analysts to make plans and manage for the emergent situations according to the types of the disasters.

The above cases demonstrate how our system supports spatial decision making through evaluation of varying-density population area to determine changes in behavior, movement, and increase overall situational assessment. This increased spatial activity and behavioral understanding provides rapid situational assessment and provides insight into evolving situational needs to provide appropriate resource allocation and other courses of action (e.g., traffic rerouting, crowd control).

We requested informal feedback for the usability of our system from users within our universities, and received useful and positive comments and suggestions. They were interested in the findings of the abnormal situations during the disaster events in Section 4.2.1 and 4.2.3. They also noted that the use of the infrastructure symbols on the heatmaps improved the legibility of the Twitter user distributions in Figure 4.2 and they suggested a visualization for the deviations between multiple heatmaps in order to show the differences clearly, which we plan to develop in the future.

4.4 Summary

We presented a visual analytics system for public behavior analysis and response planning in disaster events using social media data. We proposed multiple visualizations of

spatiotemporal analysis for disaster management and evacuation planning. For the spatial decision support, we demonstrated an analytical scheme by combining multiple spatial data sources. Our temporal analysis enables analysts to verify and examine abnormal situations. Moreover, we demonstrated an integrated visualization that allows spatial and temporal aspects within a single view. We have still some limitations with these techniques including the potential occlusion issues in the spatiotemporal visualization.

5. TRAJECTORY-BASED VISUAL ANALYTICS FOR ANOMALOUS HUMAN MOVEMENT ANALYSIS

Analysis of human movement patterns are important for urban planning, understanding the pandemic spread of diseases, disaster response, and evacuation planning in crisis management. The rapid development and increasing availability of mobile communication and location acquisition technologies allow people to add location data to existing social networks so that people share location-embedded information. For human movement analysis, such location-based social network services have been gaining attention as promising data sources. Researchers have mainly focused on finding daily activity patterns and detecting outliers. However, during crisis events, since the movement patterns are irregular, a new approach is required to analyze the movements. Also, analyzing location data alone is limited in achieving situational awareness of the events. To address these challenges, in this thesis we propose a trajectory-based visual analytics system for analyzing anomalous human movements during disasters using multi-online media. We extract trajectories from location-based social media and cluster the trajectories into sets of similar sub-trajectories in order to discover common human movement patterns. We also propose a classification model based on historical data for detecting abnormal movements using human expert interaction. In addition, we integrate multiple visual representations using relevant context extracted from different online media sources. This enhances the human movement analysis by improving situational awareness. The major contributions of this work are as follows:

- We develop a visual analytics system to discover and explore common structural movement patterns from unstructured massive movement data.
- We design a trajectory-based classification model for abnormal movement detection using human expert interaction.

- We develop visual means to improve human movement analysis using semantic context available from multiple online media sources.
- We demonstrate the effectiveness of our system in disaster management and evacuation planning through case studies.

5.1 System Overview

Our system is designed for exploring and discovering common movement patterns and detecting abnormal situations using LBSN data. The system consists of four major components: a trajectory extraction module, a data analysis module, a context extraction module, and a visualization module as illustrated in Figure 5.1. The *trajectory extraction module* (Section 5.1.1) generates two different sets of trajectories: target and normal trajectories. The *data analysis module* (Section 5.1.2) consists of two components: common movement discovery and abnormal pattern detection. For the given trajectory sets, the first component discovers major common routes based on the partition-based clustering model, and the other component assesses the abnormality for each common route. The *context extrac-*

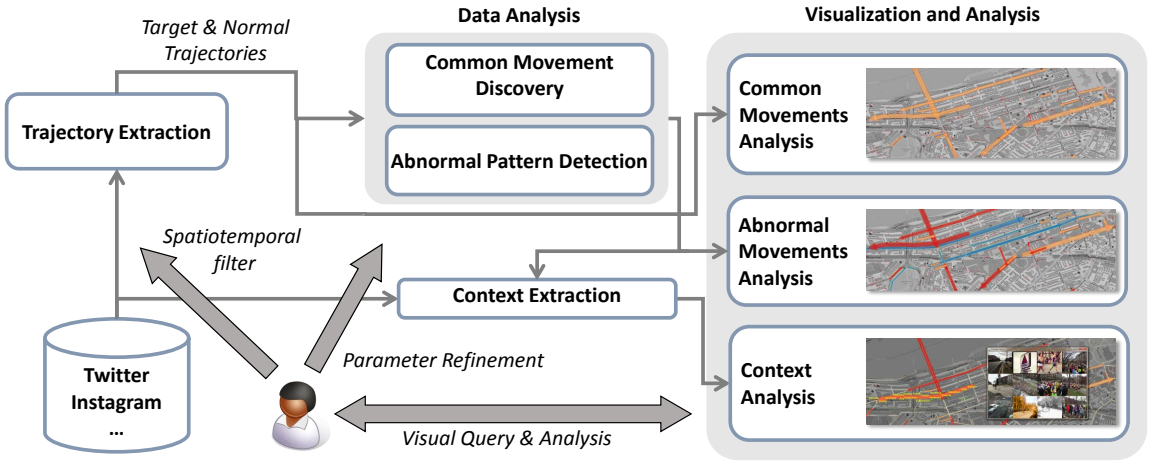

Fig. 5.1. Overview of our iterative analysis scheme for human common movement discovery and anomaly analysis.

tion module (Section 5.2) finds relevant context information including keywords, photos, videos from web cameras, and news media based on the results from the analysis module. The *visualization module* (Section 5.1.3) allows the users to explore the trajectories, common routes, and abnormal movements, and obtain a better understanding of movement patterns using additional context. Users can iteratively make visual queries and refine the parameters used for clustering and anomaly detection to optimize the results.

5.1.1 Trajectory Extraction

Our system extracts trajectories from location-based social media data. Users first select a geographical boundary and a target time window of interest. The users additionally select one or more past time windows representing normal situations to compare against the target time frame. The trajectory extraction module then requests two sets of Tweets from the database for the two selected time windows. The module generates two sets of trajectories: target and normal trajectories using geo-location information of chronologically ordered Tweets for each person.

Fig. 5.2. Supplementing a sparse trajectory (Blue) using route direction information (Yellow).

The generated trajectories, however, are usually sparse and incomplete. For example, as shown in Figure 5.2, the sparse trajectory (a blue line) does not represent a realistic movement. In order to reduce the spatial sparseness of the raw trajectories, we fill the trajectories with supplementary points between two points for each pair of the trajectories, which are calculated by shortest-path-based route directions (a yellow poly line in Figure 5.2). In

this work, we use the Bing Maps API to obtain route information. We can choose one of the following travel modes: walking, driving, or public transportation mode depending on the location and the situation. For example, we select the walking mode for the Boston Marathon case because many traffic routes along the marathon course were closed on the race day.

5.1.2 Data Analysis

In this section we describe the data analysis module. This module analyzes the trajectories given by the trajectory extraction module. The following sub-sections provide the details of this module.

Common Movement Discovery

Discovering common movements is a critical process for exploration and analysis of a large volume of trajectory data. Clustering is a popular approach in looking for common patterns in the trajectory data. Representative clustering algorithms for trajectory include DBSCAN [128] and OPTICS [129]. Andrienko et al. [28] propose a wide range of clustering-based analytics models and combine those with visualization techniques. Their clustering models, however, group similar trajectories as a whole and extract common whole trips. In this work, we utilize a modified partition-based clustering model, TRACLUS [16], in order to find common sub-trajectories. For each given trajectory, this model first partitions a trajectory into a set of line segments, and then groups the line segments into clusters of similar line segments. Clustering the line segments (as opposed to whole trajectories) eables the extraction of similar portion of trajectories. For example, Figure 5.3 shows that the three trajectories (green, black, red) have different origins and destinations, but there is a common path in all three trajectories (blue).

Clustering the line segments requires a distance function measuring the distance between line segments. We use the distance function based on a modified line segment Hausdorff distance [130], which is comprised of three components: the perpendicular distance

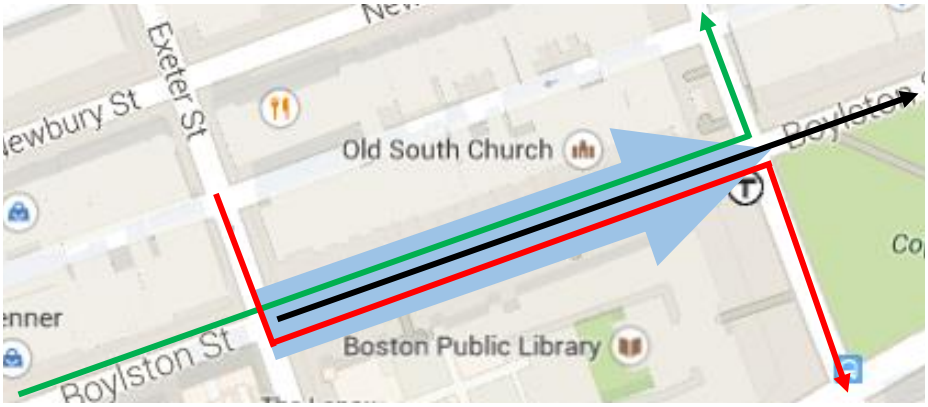

Fig. 5.3. Discovering a common sub-trajectory.

($d_\perp$), the parallel distance ($d_\parallel$), and the angle distance (d_θ). Let s_i and e_i be the starting and ending points of line L_i, and s_j and e_j for line L_j. Without loss of generality, the longer line segment is assigned to L_i, and the shorter one to L_j. These are illustrated in Figure 5.4.

For our distance function, we use $d_\perp$ and $d_\parallel$ as defined by [130], but redefine d_θ as the existing model does not consider the directions of the two line segments for the angle distance measure. In this work direction is an important factor in clustering and abnormal movement detection. To consider the direction, we utilize the cosine-similarity that measures the cosine of the angle between line segments, and is used as a bounded similarity function within $[0,1]$. $d_\theta(L_i,L_j)$ is defined as:

$$d_\theta(L_i,L_j) = \parallel L_j \parallel \times \frac{\cos^{-1}(\textit{cosine-similarity}(L_i,L_j))}{\pi} \tag{5.1}$$

where $\parallel L_j \parallel$ denote length of L_j, and θ $(0^\circ \leq \theta \leq 180^\circ)$ denote the smaller intersecting angle between L_i and L_j, and $\textit{cosine-similarity}(L_i,L_j)$ is defined as:

$$\textit{cosine-similarity}(L_i,L_j) = \cos(\theta) = \frac{\overrightarrow{s_ie_i} \cdot \overrightarrow{s_je_j}}{\parallel \overrightarrow{s_ie_i} \parallel \parallel \overrightarrow{s_je_j} \parallel} \tag{5.2}$$

$\parallel L_j \parallel$ denote length of L_j, and θ $(0^\circ \leq \theta \leq 180^\circ)$ denote the smaller intersecting angle between L_i and L_j.

The distance function is finally defined as the sum of three components:

$$\textit{dist}(L_i,L_j) = w_\perp \cdot d_\perp(L_i,L_j) + w_\parallel \cdot d_\parallel(L_i,L_j) + w_\theta \cdot d_\theta(L_i,L_j) \tag{5.3}$$

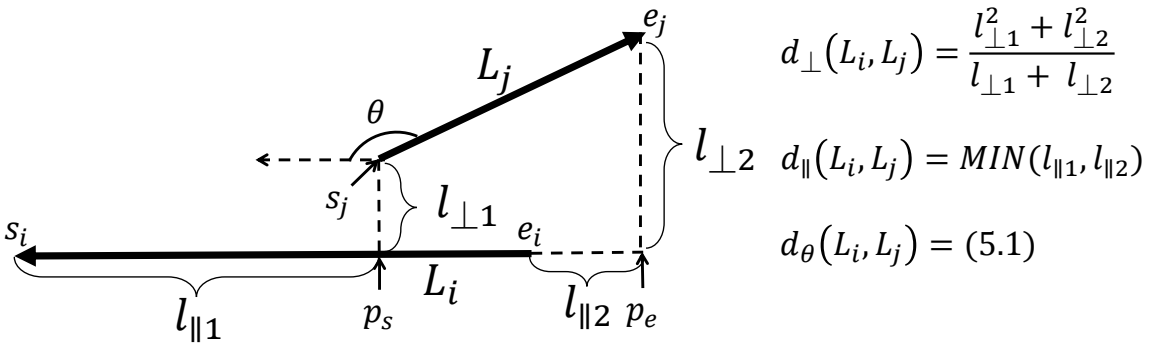

Fig. 5.4. Similarity measurement for two line segments.

where $w_{\perp}$, $w_{\parallel}$, and w_{θ} are weight values, which are determined depending on applications.

The partition-based clustering model utilized in this work is based on the algorithm DBSCAN [128]. Given a set of line segments, the algorithm groups the line segments into a set of clusters according to the distance function Equation (5.3). DBSCAN requires two parameters: ε (as neighborhood distance) and *MinLns* (as minimum cluster size). The clustering model estimates the optimal parameter values based on input data. An initial result generated by the estimated parameters is given to users. However, the automatically estimated parameter values do not always provide optimal results. Especially, *MinLns* relies on user's domain knowledge and application requirements. So, our system allows the users to manually adjust the estimated values. The system then generates a representative trajectory for each cluster, which epitomizes the line segments (sub-trajectories) belonging to the corresponding cluster. More detailed information about these procedures can be found in [16].

Abnormal Movement Detection

Existing anomaly detection models [85, 131] for trajectory data have mainly focused on identifying outliers from a target dataset. The models are usually based on non-supervised learning—they generally do not have factors for the outliers, and assume that the outliers

make for a small sub-set from the entire dataset. These models first look for major flow patterns and then determine whether each trajectory belongs to the majority according to specific criteria. However, during abnormal situations, such as natural disasters and crisis events, even the major behaviors can be unusual compared to normal situations.

In this work, we propose a classification model based on historical data for abnormal movement detection using human expert interaction. We allows the users to utilize their domain knowledge of the geographical and temporal characteristics of the location where an abnormal event of interest occurs. The users select a target time window for the abnormal situation and also chooses another time window representing a normal/regular situation for the location. We extract two sets of trajectories for the two different time frames and cluster each set of trajectories into two sets of trajectory clusters. Next, we generate two different sets of representative trajectories: target T and comparable C. A representative trajectory $RT_{ti} \in T$ is classified as an outlier if there is a close representative trajectory $RT_{ci} \in C$. More specifically, we identify outlying line segments $L_{ti} \in RT_{ti}$ [131], which are determined by the distance from neighboring RT_{ci}. We define a representative trajectory RT_{ci} is close to a line segment L_{ti} if $\sum_{L_{ci} \in CRT_c} len(L_{ci}) \leq len(L_{ti})$ where CRT_c is the set of RT_{ci}'s line segments within the distance D from L_{ti}, $\{L_{ci} \mid dist(L_{ci}, L_{ti}) \leq D\}$. A larger value of D detects a smaller number of outliers, and a smaller value of D a larger number of outliers. Then, intuitively a representative trajectory RT_{ti} is outlying, if the percentage of the total length of outlying line segments is more than P. The default value of P is set to 30. Finally, the outlying representative trajectories are visualized. Our system allows the users to adjust the two parameter values, D and P in order to refine their results.

5.1.3 Visualization and Analysis

In this section we describe our design goals. We introduce the visualization module to show common and abnormal movement patterns discovered by the data analysis module. To illustrate our method, we use the Tweets generated near the finish line of the Boston Marathon during first 24 hours after the two explosions on April 15, 2013.

Design Rationale

Our design goal is to show trajectories extracted from geo-tagged Tweets of each person. Displaying the trajectories without grouping can also reveal new insights when users drill down to individual movements. However, when the number of trajectories shown over the map increases, visual clutter issues arise that hinder the discover of flow patterns. To reduce clutter, we use a modified partition-based trajectory clustering model for discovering common sub-trajectory patterns [16]. The discovered common movements have multiple attributes to be analyzed, such as cluster size, direction, and length. The users need to not only identify abnormal movement patterns, but also understand how abnormal and normal movement patterns differ. The required clustering level can also vary with the application. So, we need to allow the users to adjust the clustering level.

Visualization of Common Movements

Figure 5.5 shows the process of discovering common movement patterns. If we display the raw trajectories, it is difficult to understand the realistic movement patterns because of the high degree of sparseness of the trajectories as shown in Figure 5.5 (left). Therefore, we reduce the sparseness of the raw trajectories using the method described in Section 5.1.1 and display the supplemented trajectory on the map with 30% opacity in Figure 5.5 (center). Users are able to examine more realistic human mobilities with the supplemented trajectories rather than using the raw ones. Next, we cluster the trajectories into sets of similar sub-trajectories and generate representative trajectories for each cluster as described in Section 5.1.2. The representative trajectories represent common movement behaviors in Figure 5.5 (right).

We provide visual cues to show multiple attributes for a representative trajectory. We use a poly line with an arrow head to display the length and the direction of the representative trajectory. The thickness of the line represents the size (i.e., the number of sub-trajectories belonging to a cluster) of the corresponding cluster. Figure 5.6 shows the representative trajectories within the region same as the one in Figure 5.5 (right). Placement

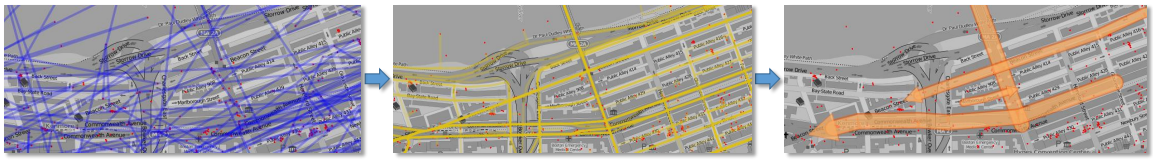

Fig. 5.5. The process of discovering common human movement patterns using location-based social networks data. Visualization of sub-trajectory clusters (right). The thickness of each trajectory represents the size of the cluster.

order of the trajectories depends on the length; the longest trajectory is placed at the bottom and the shortest one at the top, to avoid obscuring the shorter trajectories.

Our system also enables users to adjust and refine the ε (as neighborhood distance) and *MinLns* (as minimum cluster size) values used by the clustering algorithm depending on their requirements by visual inspection. We display an initial clustering result calculated with the automatically estimated parameter values as described in Section 5.1.2. The optimal result in Figure 5.7 (top) is achieved at $\varepsilon = 25$ and $MinLns = 3$. The algorithm generates a larger number of smaller clusters, when ε is smaller or *MinLns* is larger compared to the optimal values. In contrast, the algorithm generates a smaller number of larger clusters when ε is larger or *MinLns* is smaller. For example, the result at $\varepsilon = 25$ and $MinLns = 4$ is shown in Figure 5.7 (center) and the results at $\varepsilon = 30$ and $MinLns = 2$ is shown in Figure 5.7 (bottom).

Visualization of Abnormal Movements

Our analytics model identifies abnormal representative trajectories from target ones by comparing with normal ones as described in Section 5.1.2. We define target outliers are the abnormal representative trajectories and target normal trajectories are the rest of the target representative trajectories; the target normal trajectories are close to the normal representative trajectories. We visualize the three different types of representative trajectories: target outlier, target normal, and normal using a similar visual encoding scheme described in the

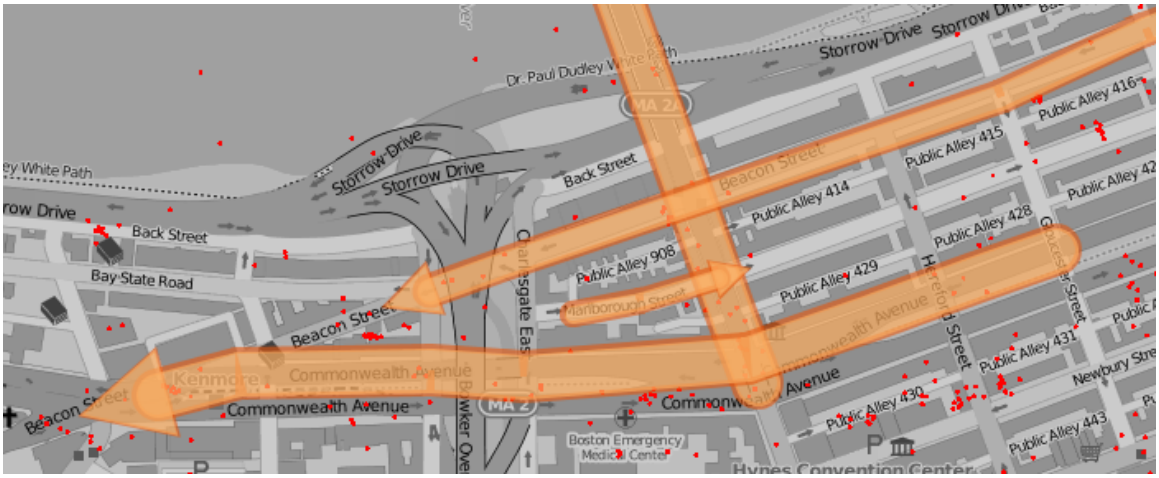

Fig. 5.6. Visualization of sub-trajectory clusters. The thickness of each trajectory represents the size of the cluster.

previous section. We use different colors to distinguish between the different types: target outlier with red, target normal with orange, and normal with blue as shown in Figure 5.8. We can see that the trajectories (1), (2), and (3) are close to the normal trajectory (4), but they head toward the opposite direction. Those are, therefore, classified as outliers. The trajectory (6) is not classified as an outlier, because it is close to the normal trajectory (5) and also has the same direction. We also provide an option to turn on/off each type to focus on a specific type.

5.2 Improving Analysis using Multi-Context Information

In this section we describe our design goals of usage of each context. We describe how we extract the context from multiple data sources. Also, we show how the context is visually incorporated into the system.

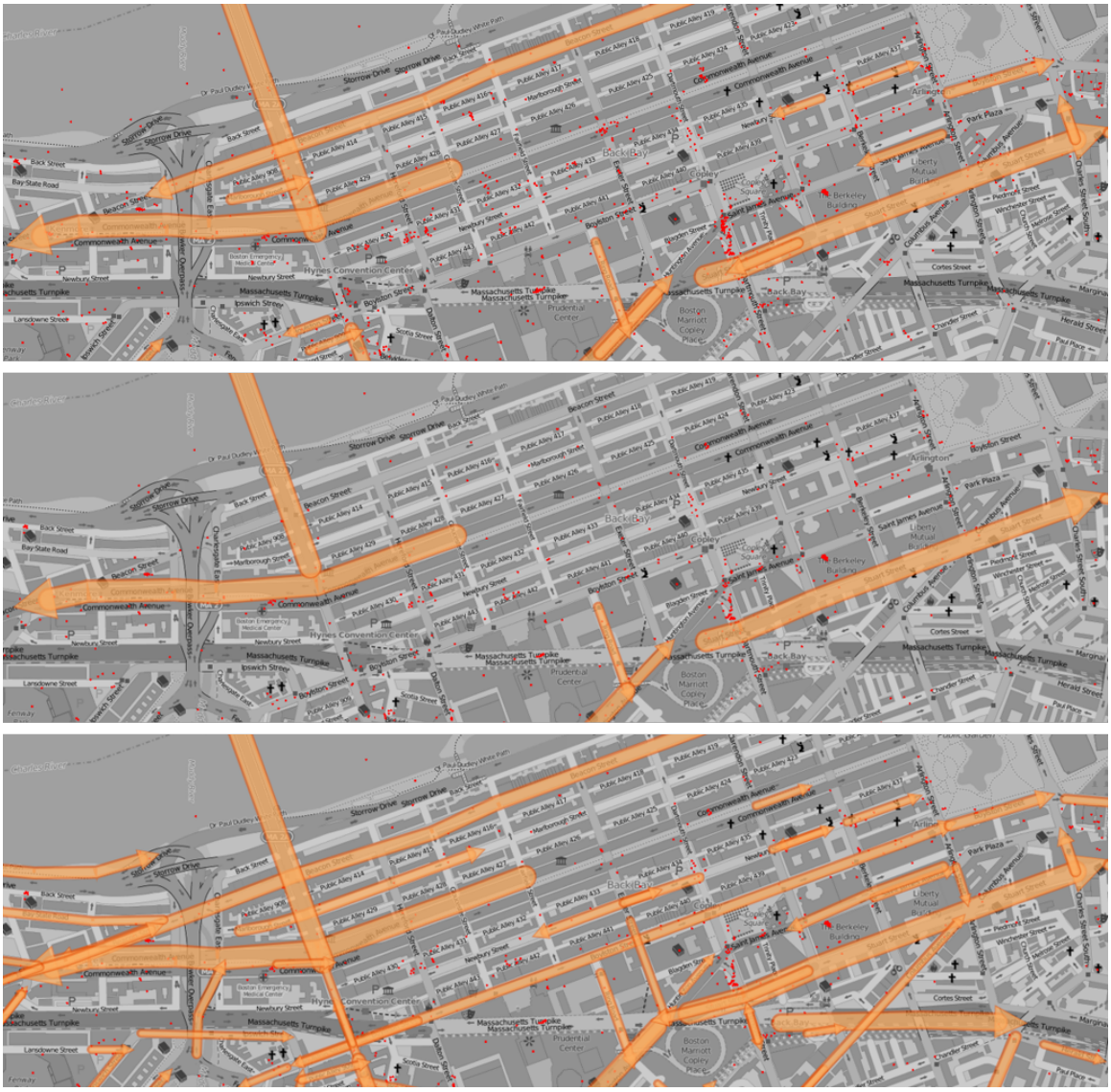

Fig. 5.7. Clustering results depending on two parameters: ε and *MinLns*. Top ($\varepsilon = 25$, $MinLns = 3$) is optimal. Center ($\varepsilon = 25$, $MinLns = 4$) shows less number of trajectory clusters. Bottom ($\varepsilon = 30$, $MinLns = 2$) shows more.

5.2.1 Keyword Extraction and Visualization

Analyzing the spatial behaviors alone is limited in achieving situational awareness of local events—they cannot answer why people move and what situations occur. To address

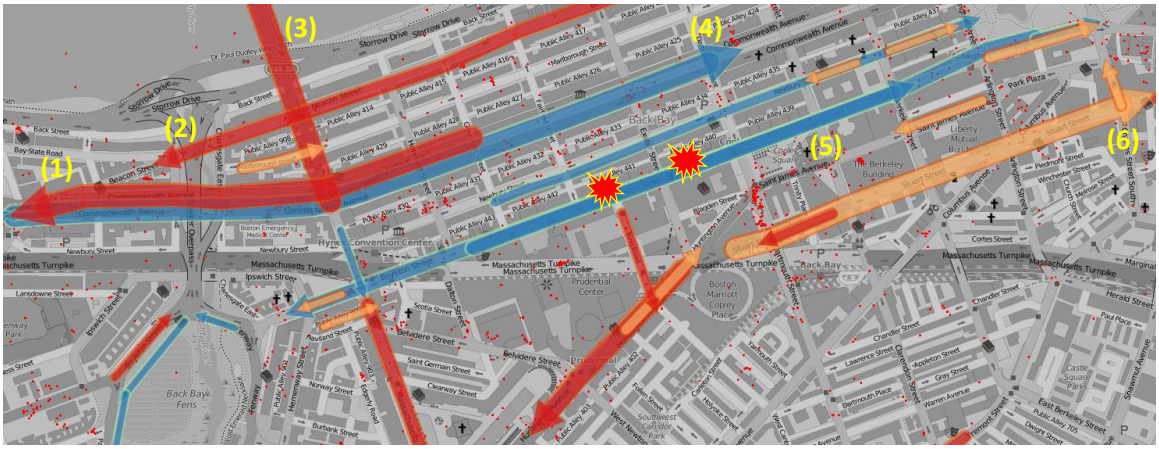

Fig. 5.8. The trajectories (red and orange) shows the human movement patterns around the finish line at the Boston Marathon 2013 during 2 hours after the explosions. The trajectories (blue) represent the movements for the normal situation (the same time period of the same event in 2014). The two markers indicate the locations of the two explosions.

this challenge, we extract keywords from the tweets used to generate a trajectory cluster and also those located close to the cluster, because such Tweets can contain common topics indicating an event occurring around a similar mobility. Then we display the extracted keywords for providing additional insights into the event and the mobility patterns.

We select a set of Tweets that constitute sub-trajectories belonging to a cluster and are located within a specific distance to the representative trajectory of the cluster. In this work, the default value of the distance threshold is set to 200 meters. Then, we extract keywords from the text of the selected Tweets. We calculate the frequencies of each word in the aggregate text and select top keywords based on the frequencies.

To display the extracted keywords, we utilize 'tag cloud'visualization. Tag clouds have been used to represent a most frequent or important words in order to summarize text collections [132]. Also, tag clouds can be exploited for analytics tasks, such as topic-based document navigation and labeling geographical points of interest [75, 133]. In this work, we display the keywords along their corresponding representative trajectory without overlapping. The font size of each keyword encodes the frequency to show the popularity

of the keyword. Figure 5.9 (top) shows an example of showing extracted keywords display along the center trajectory.

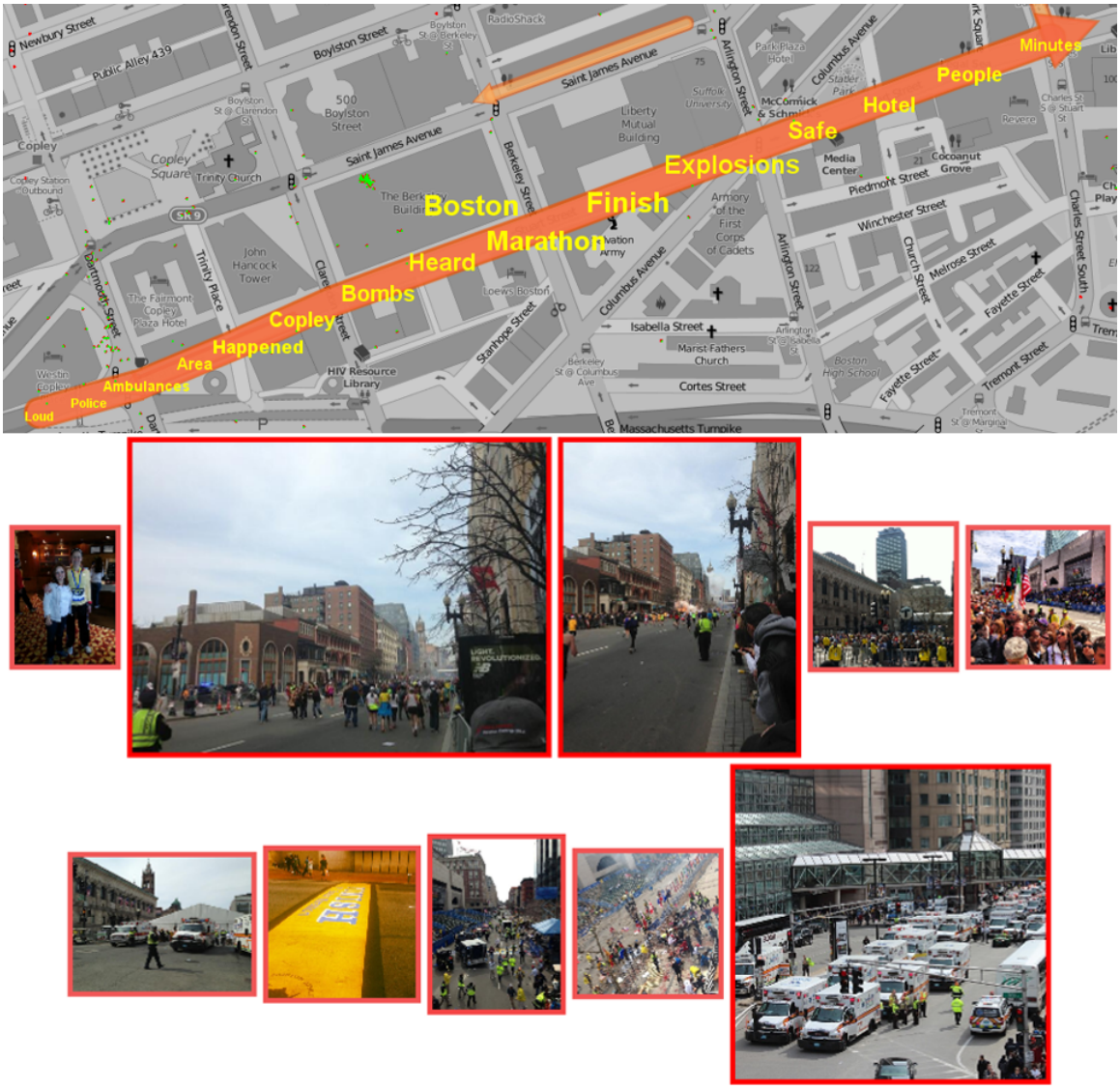

Fig. 5.9. The extracted keywords along the trajectory close to the explosion locations show a strong relationship to the explosions (top). The chronologically displayed photos (bottom) extracted from the same trajectory show the scenes of evolving situations.

5.2.2 Additional Context Information

For additional context, we utilize shared photos, public web camera videos, and news media related to each representative trajectory. These data sources provide additional channels of information to users; thereby, providing a more comprehensive situational awareness for any emerging situation.

Shared photos: Right after people take photos, they can post the photos to LBSNs with their smartphones. For example, we collected more than 230 of Tweets with photos were generated within the first 5 hours after the explosions at the Boston Marathon in 2013. The photos allow first responders and emergency managers to obtain a better situational awareness of what is happening on the site during a crisis.

In our system, we first select the Tweets in the same manner as for keyword extraction. we identify Tweets with photos from the selected tweets for each trajectory cluster; tweets with photos contain photo links in a particular format. Images are retrieved from links within Tweets and displayed chronologically in a separate window (e.g., as shown in Figure 5.9). Photo sizes correspond to relative differences in their sharing count (sum of retweets and replies) [134]. When a photo is selected, the location of the photo's Tweet is highlighted on the map.

Public live web camera videos: We utilize public webcam video feeds to allow the emergency managers to obtain a better situational awareness of an emerging crises situation [135]. In our system, we mark available camera locations on the map with pre-loaded camera location data [136]. Once users click on one of the web cameras icons on the map, the live streaming feed or most recent snapshots are provided (e.g., as shown in Figure 5.10).

News media: We augment the information provided by the public in social media with news media reports from major news agencies. News media reports related to the context of movements can provide more reliable information. We search news reports with extracted keywords from nearby Tweets for each cluster using the Google search APIs.

The results, including titles, summaries and links of each news report, are displayed in a separate window.

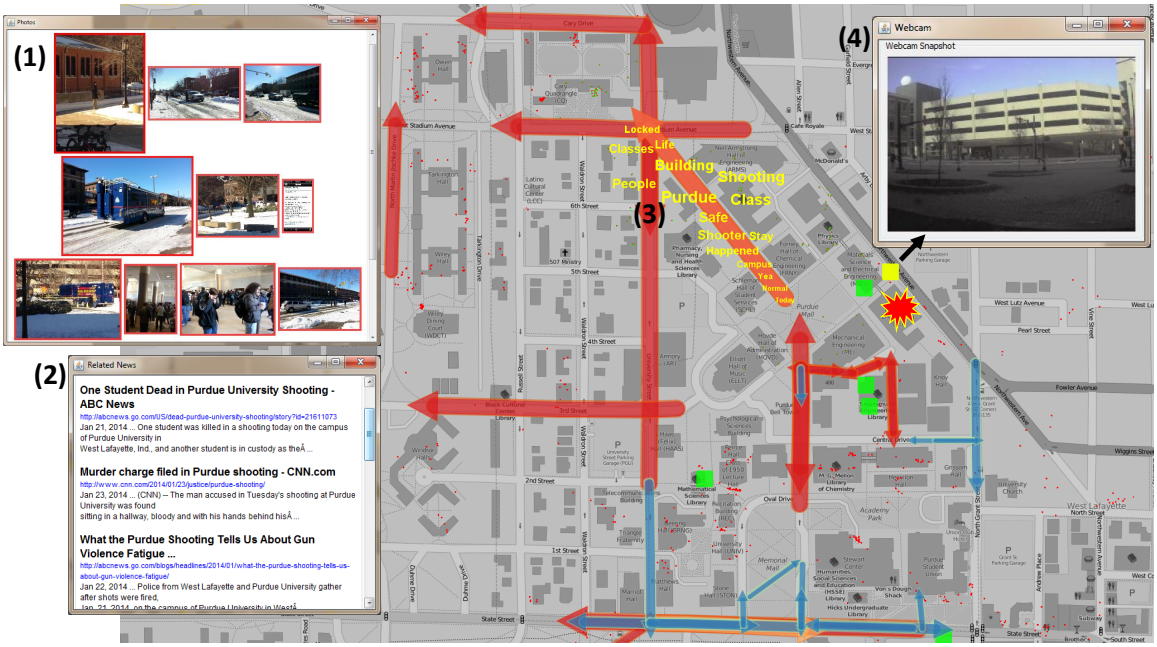

Fig. 5.10. The trajectories (red and orange) shows the human movements around the campus during 2 hours after the shooting. The normal trajectories (blue) extracted from the same time period on normal day. Photos (1), News reports (2), Keywords (3), and Webcam videos (4). The green rectangles indicate the locations of the web cameras around the campus. The yellow one is the selected camera. The marker indicate the building where the accident occurred.

5.3 Case Study

In this section we demonstrate how our analytics model can assist emergency managers in discovering common/anomalous human movement patterns during crisis events, and how our visual analytics system improves movement analysis for disaster management personnel.

5.3.1 Boston Marathon Explosion

Boston Marathon is an annual marathon held in Greater Boston and one of the world's best-known athletic events. On April 15, 2013, two bombs exploded near the finish line during the Boston Marathon at 2:49 pm EDT. Figure 5.8 shows two markers that indicate the locations of explosions. The trajectories in Figure 5.8 show the movement patterns at the Boston Marathon, where the orange colored trajectories show the movements during the Boston Marathon bombing using Twitter data for the 2 hours after the explosions The blue colored trajectories represent the normal movements using the data from next years' Boston Marathon event (we use next year's data for illustrative purposes instead of the previous year's data due to the unavailability of data for the previous year in our database). The system utilizes these two trajectories in order to compute the abnormal trajectories (shown in red). The target trajectories (shown in orange) show that people were dispersed from the locations of the explosions and did not use the road where the accidents occurred. Also, the outlier trajectories 1, 2, and 3 in Figure 5.8 show that participants and spectators moved in the opposite direction of the finish line or crossed the bridge in order to get away from the location of impact. Furthermore, Figure 5.9 (top) shows the keywords and photos extracted along the trajectory labeled 6 in Figure 5.8 (note that the the photos chronologically displayed in the system). Since the trajectory is close to the explosion locations, the extracted keywords along the trajectory show a strong relationship to the accident. The system can thus enable first responders and law enforcement to detect anomalous movements and maintain a situational awareness of an emerging situation in their areas of responsibility.

5.3.2 Purdue University Shooting

On Tuesday, Jan 21st 2014, a shooting occurred inside one of the buildings of Purdue University, Indiana (shown by the marker in Figure 5.10). Figure 5.10 shows the movement patterns of people around the campus during 2 hours after the incident, where red colored trajectories show anomalous behavior, and orange colored trajectories show the movements

during the 2 hours after the incident. We compare the movements to the normal movements (blue) extracted from the same time period on another Tuesday. We can observe anomalous behavior from the results where people moved to the left or upper-left regions. Upon further investigation, we find that these locations house student residence halls. Only a few people moved around the site of the incident, because of a lock down order given by the police. In Figure 5.10, the photos (1) provide a visual context extracted from nearby Tweets of the trajectory (i.e., the scenes around the area and inside buildings). The keywords (3) extracted from the selected trajectory convey more information describing the accident. The news reports (2) extracted using the keywords along the tracectory (3) allow users to get more detailed information about the event. Finally, the video feed (4) enables users to monitor the region in real time. Emergency managers can thus utilize social media as another input information source to maintain a situational awareness using our system.

5.4 Summary

We presented a trajectory-based visual analytics system, making it possible to: 1) generate trajectories using geo-tagged Tweets, 2) discover human common movement patterns, 3) detect abnormal movements, and 4) improve human movement analysis using semantic context available from multiple online media sources. In order to find common movements, we utilize an enhanced partition-based clustering model that allows to extract similar portion of movements. We proposed a classification model using human expert interaction to identify abnormal movements. We described how we effectively extract and utilize relevant context, such as keywords extracted from Tweet text, shared photos, web camera videos, and news media for providing a better understanding of spatial movement behaviors. We demonstrated the usage and effectiveness of our system for human movement analysis in abnormal situations by case studies.

6. FORECASTING THE FLOW OF HUMAN CROWDS

Researchers from various domains have put considerable effort into modeling the mobility of individuals to understand their movement patterns using different data sources. A wide range of applications, such as urban planning and traffic planning, depend greatly on the movement behaviors of large crowds. In this thesis, we introduce a novel visual analytics approach for forecasting the overall flow of these human crowds. Given a space with a large number of moving individuals, our model partitions the space into smaller sub-spaces, and then calculate the directional density of flows for each sub-space. We apply seasonal trend analysis techniques on the directional density data in each sub-space to forecast the future crowd movement based on the observed historical flow patterns. We then combine the predicted results to visualize the overall future flow. Our methodology considers road directions for more accurate directional flow density estimation and applies a data imputation technique to mitigate data sparsity issues. We present results from a series of statistical tests for evaluation of our methodology across different spatial movement datasets (e.g., location-based social network, GPS tracks of humans and taxis). The main contributions of this work include the following:

- We introduce a new method to estimate the directional flow density that represents the overall movement directions of moving objects over a 2D space.
- We propose a new model to forecast the flow of human crowds based on the historical directional density data.
- We develop a new flow visualization technique of multi-vector fields to represent the directional flow density.

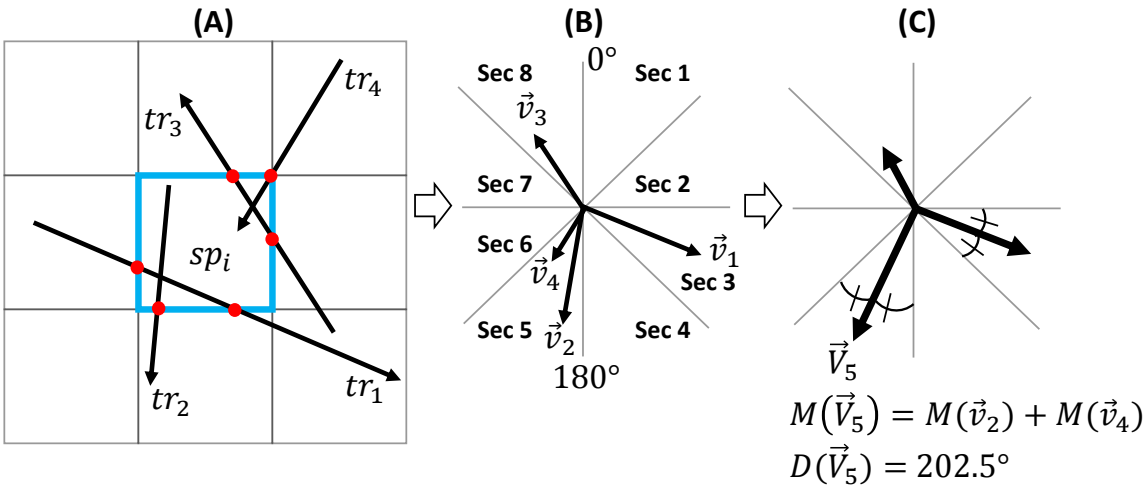

Fig. 6.1. Trajectory tessellation and directional density estimation

6.1 Flow Data Modeling, Forecasting, and Visualization

Our flow modeling methodology consists of a pipeline of several processes. We first apply a directional density estimation technique (Section 6.1.1). As the next step, we provide two different procedure choices: flow smoothing (Section 6.1.3) and missing-data imputation (Section 6.1.5), where the two procedures is to mitigate the data sparseness. After these processes are completed, we forecast the future flows using the seasonal trend decomposition based on loess technique (Section 6.1.4) and finally visualize the results using multiple flow visualization techniques (Section 6.1.6). These processes are described in detail in the following sub-sections.

6.1.1 Directional Density Estimation: Classification of Fixed Direction Sector

Given a set of trajectories $TR = \{tr_1, \cdots, tr_{num-tra}\}$ for a specific time window t, we equally divide a space $\Omega \subset \mathbb{R}^2$ into smaller sub-spaces $\Omega = \{sp_1, \cdots, sp_{num-spa}\}$. We then estimate the directional density for each sub-space that represents the overall movement direction for each sub-space using the following methodology (Figure 6.1). For each sub-

space sp_i, we first segment the individual trajectories tr_i that occur in each sub-space grid by checking the crossing points of the trajectory segments at the boundary. Figure 6.1 (A) shows the result of this process where six points (highlighted in red) are detected and are assigned to be either the start or end locations for each segment that passes through the grid (highlighted in blue). These points are used to identify partial trajectories (i.e., sub-trajectories) for a trajectory tr_i. Next, we transform the sub-trajectories into a Euclidean vector space $\vec{v}_i$ and translate the start points and end points in the space to align with the center location of the sub-space. That is, the starting points of the sub-trajectories within the sub-space are located at the center of the sub-space (Figure 6.1 (B)).

The next step in our approach is to summarize these vectors in order to generate meaningful representative movement vectors for each sub-space. The conventional approach of summarizing vectors includes performing computations (e.g., average, addition) over the entire space. However, this approach is often not optimal in generating representative movement vectors as the final summary vectors could be meaningless. For example, if we have two same magnitude but opposite direction vectors, they will cancel each other out to yield a zero resultant vector. Accordingly, there is a need to preserve meaningful vectors after summarization. Our approach is designed in order to help preserve the original directions of the vectors. For each sub-space, we divide one full turn (360°) into S circular sectors of the same size. For demonstration, we use 8 sectors (i.e., $S = 8$) as default configuration in Figure 6.1 (B), where each sector covers a 45°region. For each sector k (where $k = 1, \cdots, S$), we generate a representative vector $\vec{V}_k$ by aggregating the corresponding vectors $\vec{v}_i$ within the sector k (as demonstrated in Figure 6.1 (C)). The magnitude $M(\vec{V}_k)$ of the representative vector for each sector is the sum of magnitudes of the vectors ($M(\vec{v})$) that belong to the corresponding sector. The direction $D(\vec{V}_k)$ of the representative vector for each sector is the angle calculated from the north. In this way, each sub-space will have a set of S representative vectors $\mathscr{R}_{sp_i} = \left\{\vec{V}_1, \cdots, \vec{V}_S\right\}$. The representative vectors encode the directional density of a sub-space, and summarize the directions of flow for each individual sub-space.

6.1.2 Directional Density Estimation: Considering Road Direction

In addition to the fixed direction classification method described in Section 6.1.1, we propose another method considering road direction. Using the fixed direction sectors to classify flow directions has a limitation in representing actual flow directions. The method can make large distortion of moving directions because it uses fixed same directions for every sub-space, even if it can have different road directions. Eventually, it can cause inaccurate directional density of flows over a space. Thus, to resolve this issue, we propose another method that classify flows directions based on road directions. Figure 6.2 shows the calculation process of the new method. Given the raw trajectories passing over a space as shown in Figure 6.2 (A), we first calculate the shortest path of each trajectory, which reflect the road directions in Figure 6.2 (B). Then, for each sub-space, we segment the individual trajectories by checking the crossing points of the trajectory segments at the boundary. These points are used to identify sub-trajectories of a whole trajectory. Next, we transform the sub-trajectories into a vector space and translate the start and end points align with the center location of the sub-space in Figure 6.2 (C). And, accumulate the vectors according to their directions . We compute this estimation for every sub-space and generate a specific vector field. We call it as multi-vector field, because traditionally each location has only one vector, but for our case, each location can have multiple vectors. For the new method the directions are adapted to the actual road directions of each sub-space so that the directional density estimation become more accurate. In Figure 6.2 (C), the purple dashed arrows are the result by using fixed direction sectors. In addition, the new method provides another benefit. The number of directions to be computed decreases as the sub-spaces have 4 directions or less in most cases, while the fixed direction sector method have to consider at least 8 or more directions for every sub-space.

We visualize the directional densities by a glyph-based visualization as shown in Figure 6.3. The directions and the lengths of the blue arrows in a sub-space represent the directions of the flows across the sub-space and their magnitudes, respectively. Figure 6.3 shows two different results of directional densities around Manhattan in New York City.

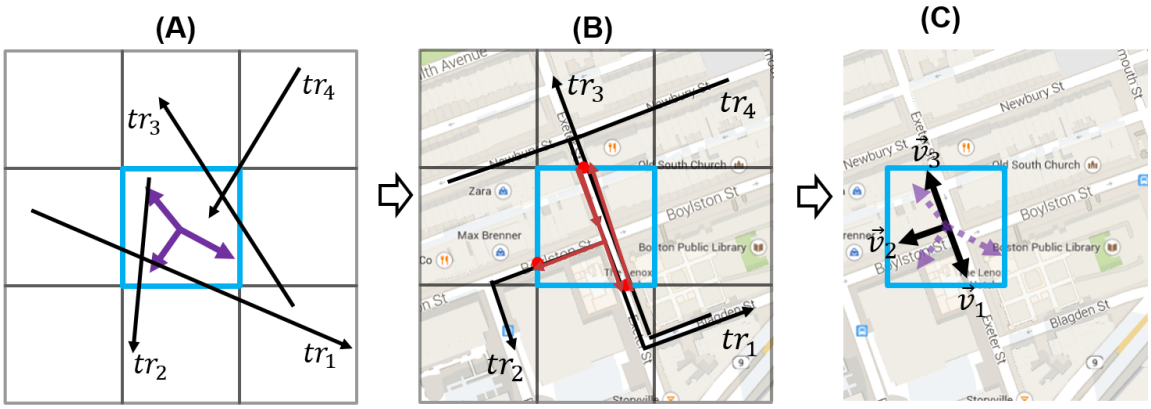

Fig. 6.2. Trajectory tessellation and directional density estimation based on shortest path considering road directions

The left result is calculated by using fixed direction sectors and the right one by the road direction method. In Figure 6.3 the yellow lines are the raw trajectories (Left) and the shortest paths (Right) calculated based on the raw trajectories. In result, the directional densities shown in Figure 6.3 (Right) indicate more accurate and realistic human mobilities. Figure 6.4 shows the directional densities around the downtown area in Chicago. For even this case, the directional densities in Figure 6.4 (Right) reflect more realistic movement paths of people than the results in Figure 6.4 (Left).

6.1.3 Flow Smoothing

Sparse and noisy flow data often generate non-smooth flow patterns that cannot be used for accurate prediction. In order to mitigate the effect caused by the data sparseness and noise, we propose a new flow smoothing method based on local and global trend estimation. The rationale behind our algorithm is that individuals in a crowd tend to follow dominant paths of the crowd [137].

In our algorithms, for each sub-space, we adjust directional density with consideration of neighbor sub-spaces' trends and a global movement that considers the entire space. First, local neighbor sub-spaces of sp_i is defined by $N(sp_i) = \{sp_j \in \Omega \mid sp_j \; is \; adjacent \; to \; sp_i\}$

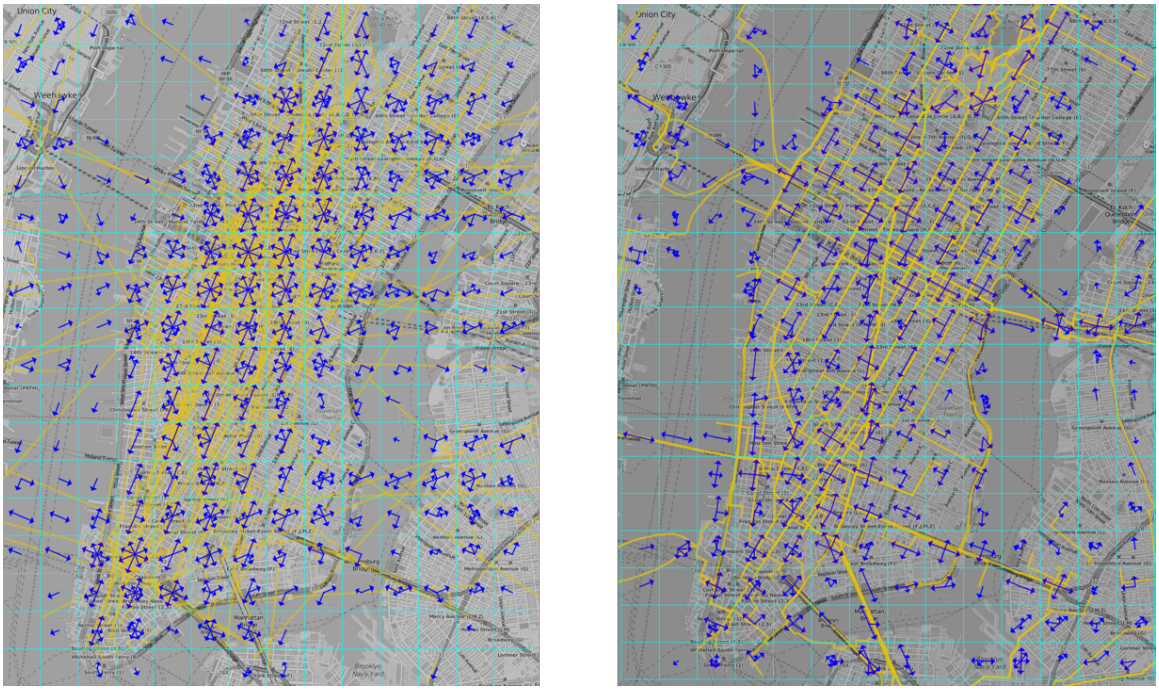

Fig. 6.3. Directional density of New York City resulted by the method 1 (Left) and the method 2 (Right)

and the local directional trend of sp_i is computed using the neighbor representative vectors $\mathscr{R}_{sp_j}, sp_j \in N(sp_i)$.

Next, we define the global movement of a sub-space as a major movement of a larger space including the sub-space. In order to compute a global trend, we extract the major movements from the trajectories of the entire space using a density-based trajectory clustering algorithm [16]. Figure 6.5 (A) shows an example computation process. Here, the representative vectors $\mathscr{R}_{centerSpace}$ of the center sub-space are almost evenly distributed in multiple directions before smoothing but the major local and global trends move toward the bottom-right. Thus, after smoothing based on the trends, the vector $V_4 \in \mathscr{R}_{centerSpace}$ heading toward the bottom-right becomes a major vector.

The next question is that which sub-spaces should be affected by the global trend since there is uncertainty in the influence of a global trend. Therefore, we assume that the sub-

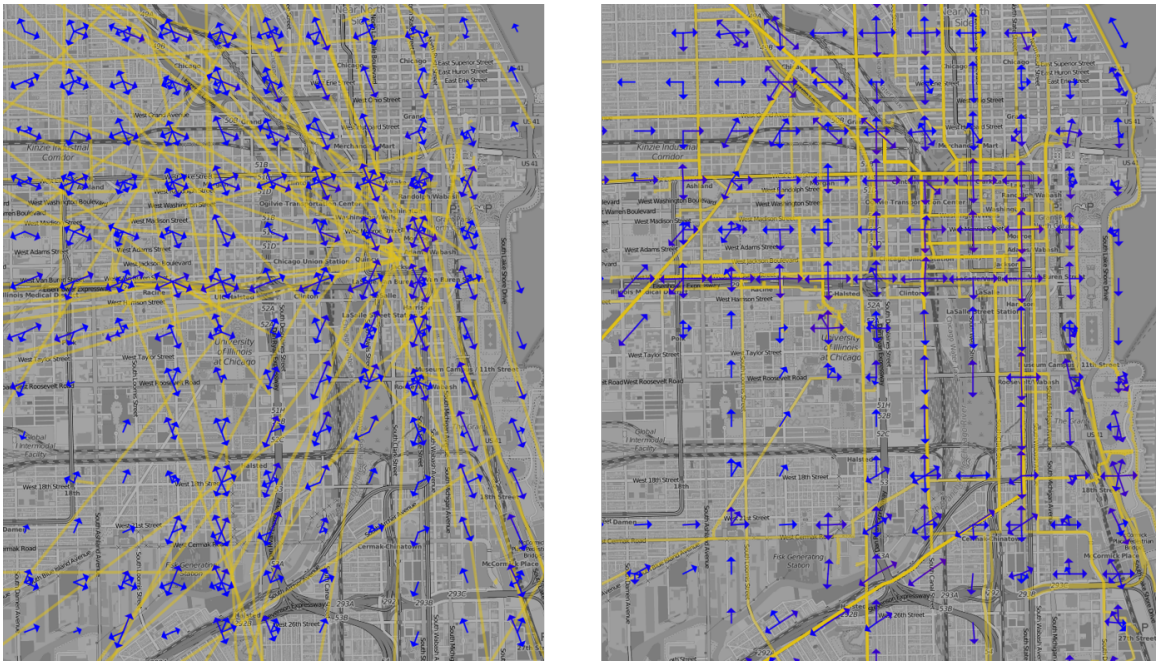

Fig. 6.4. Directional density of New York City resulted by the method 1 (Left) and the method 2 (Right)

spaces that are close to the global trend are more affected than the ones more distant. Figure 6.5 (B) shows how we classify the sub-spaces where dark gray sub-spaces are more influenced by the global trend than the light gray ones that are the neighbors of the dark gray ones. The white sub-spaces are not within the influence area.

As shown in Algorithm 1, all sub-spaces are visited in our algorithm to perform individual smoothing operations (line 1, 2). In each visit, the average magnitude is computed with consideration of the neighbor sub-spaces (line 3-12). Then, our algorithm computes interpolation based on original magnitude of sp_i, the average magnitude of neighbor sub-spaces, and the global magnitude based on a local smoothing parameter λ and a global smoothing parameter τ (line 13-18). Finally, the algorithm updates the magnitude of each sector for each sub-space based on the interpolation results.

Figure 6.6 shows the local and the global directional densities trends of taxi movements with our glyph-based visualization in southern Manhattan during in the morning on

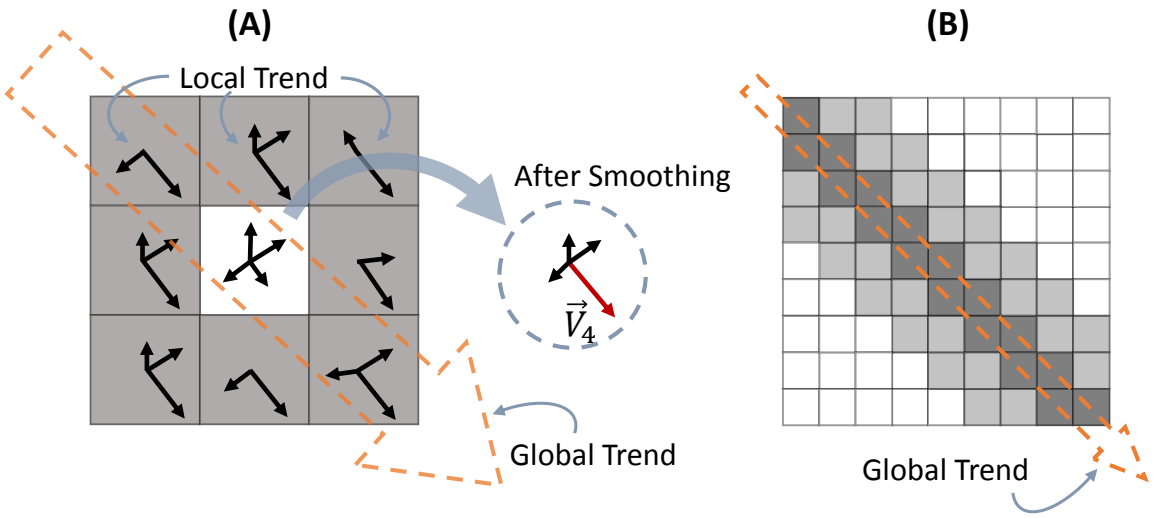

Fig. 6.5. Each space's directional density is computed based on its neighbors' and global trends.

May 24th, 2013. Each sub-space has a set of arrows where each arrow indicates the corresponding representative vector. The orange long arrow represents the global movement in the area. The yellow lines are actual taxi trajectories. The local directional densities and the global movement move toward bottom-left, as the observed time window is in the morning. Figure 6.7 shows an example smoothing result where the directional density of the sub-space A changes along the local directional trend. Note that before the smoothing process is performed, the sub-space B did not have the directional density due to data sparseness. After smoothing, it has one accommodating the local directional trend.

6.1.4 Forecasting Future Flow

To forecast directional density using historical vector-based crowd data, we apply seasonal trend decomposition based on loess technique [3] over the entire space. Figure 6.8 illustrates the overall process of our forecasting method. For a given geo-location boundary Ω and a past time window t, we first prepare the geospace with the trajectory data (Figure 6.8 (A)). Next, we fragment the geospace Ω into sub-spaces and compute a set of vec-

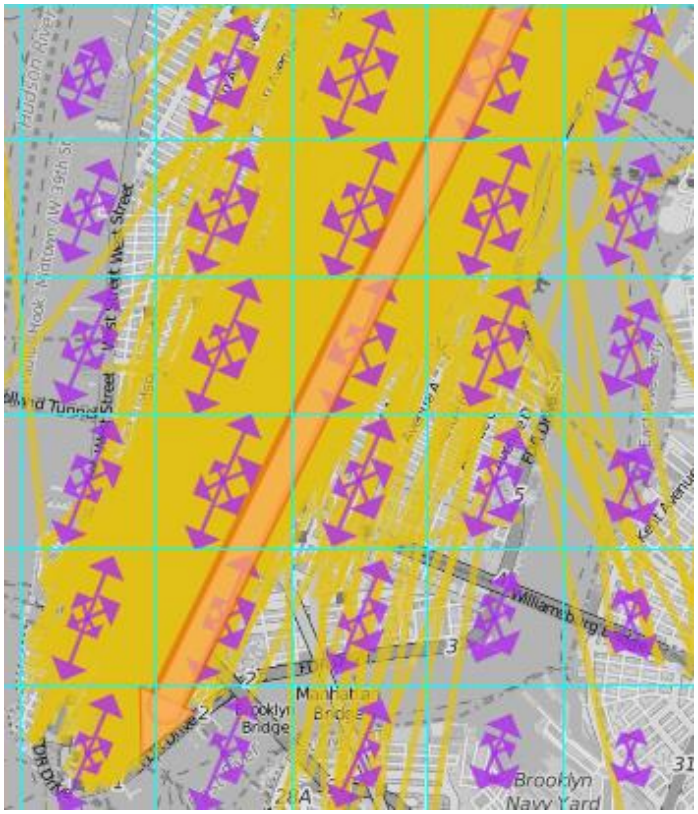

Fig. 6.6. Local directional density (purple arrows) and global (orange arrow) trends of taxi movements in southern Manhattan between 6:00 AM and 10:00 AM.

tors for each individual sub-space using the methodology described in Sections 6.1.2. As discussed previously, we generate a specific vector field $\vec{V}_k$ for every sub-space. We call it as multi-vector field, because traditionally each location has only one vector, but for our case, each location can have multiple vectors. This is shown in Figure 6.8 (A). In our approach, we define these vector fields for the given time window t as $VF_t = \left\{\mathscr{M}_{sp_i},\ \ sp_i \in \Omega\right\}$, where $\mathscr{M}_{sp_i}$ is a set of vectors for sub-space sp_i. This process is repeated for every sub-space in geospace for a given time step, and then over time for the entire time window t (Figure 6.8 (B)). Next, we generate a time series of the *magnitude* values of the vectors of the series of $\mathscr{M}_{sp_i}$ of the sub-space (Figure 6.8 (C)). This time series is defined as:

$$Y_k = \left\{M(\vec{V}_k) : \vec{V}_k \in \mathscr{M}_{sp_i}, \mathscr{M}_{sp_i} \in VF_t, t \in T\right\}, k = \{1, \cdots, S\} \tag{6.1}$$

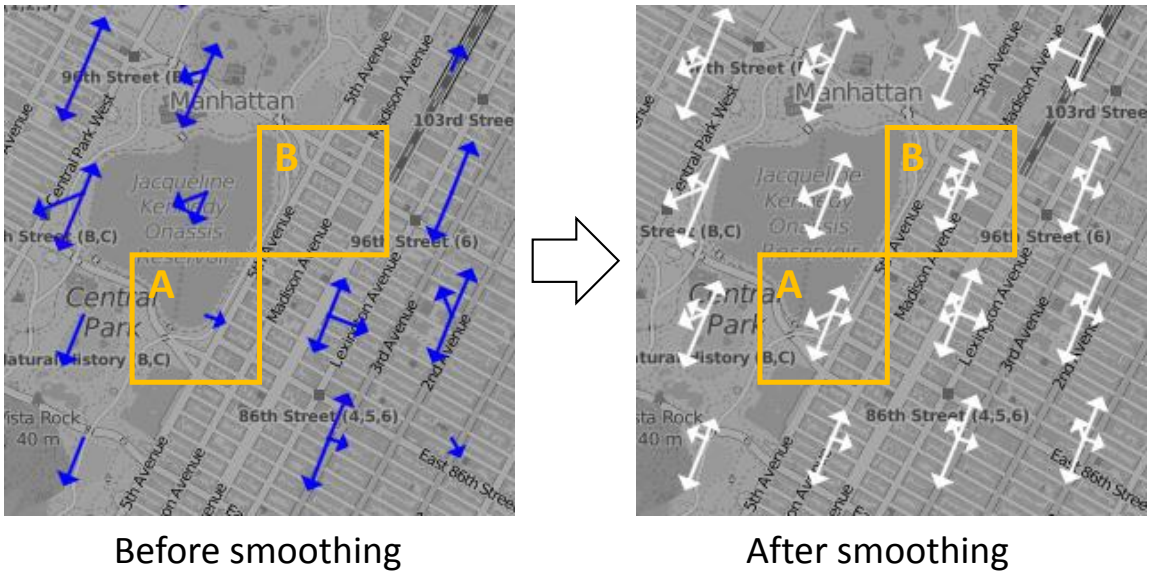

Fig. 6.7. An example of the smoothing result. Sub-space B does not have density due to sparseness data (Left). Sub-space has the density (Right).

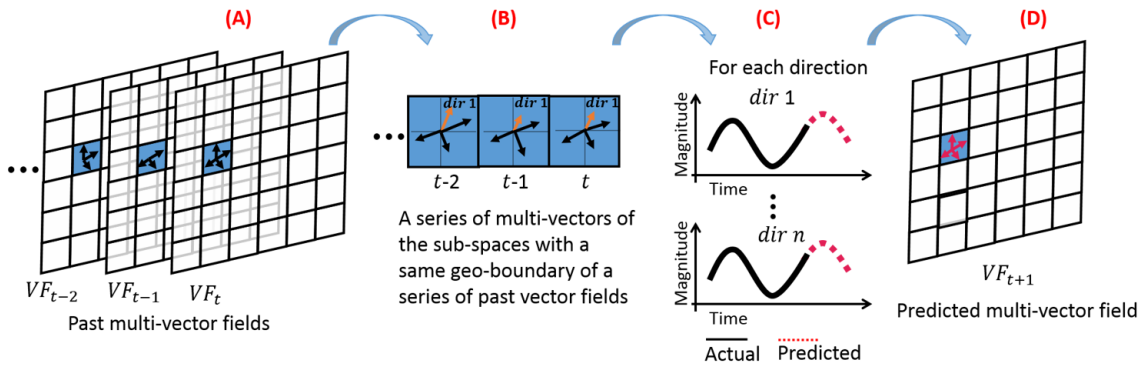

Fig. 6.8. The process of multi-vector field prediction.

Here, T is the time range of observed historical data.

In order to model the time series Y_k and forecast for the future value, we employ the seasonal-trend decomposition technique (STL) described in [138, 139]. The technique is based on a locally weighted regression (loess) methodology (STL) [3]. For each sub-space,

Algorithm 1: Flow Smoothing

Input : (1) $\Omega = \{sp_1, \cdots, sp_n\}$

(2) Global vector of sp_i G_{sp_i}

(3) A local smoothing parameter λ

(4) A global smoothing parameter τ

Note: $(0 \leq \lambda + \tau \leq 1)$

Output: Smoothed flow of each sub-space

`/* For each sub-space, smooth its directional density */`

for ***each*** $sp_i \in \Omega$ **do**

  `/* For each sector, adjust its magnitude based on the local and global directional trends */`

  for $k = 1$ *to* S **do**

    let $avgNeighborMag = 0$

    let $numNeighbor = 0$

    `/* For each neighbor sub-space of` sp_i `*/`

    for ***each*** $sp_j \in N(sp_i)$ **do**

      $m = M(V_k), V_k \in \mathscr{R}_{sp_j}$

      if $(m > 0)$ **then**

        $avgNeighborMag = avgNeighborMag + m$

        $numNeighbor = numNeighbor + 1$

      end

    end

    $avgNeighborMag = avgNeighborMag / numNeighbor$

    $originalMag = M(V_k), V_k \in \mathscr{R}_{sp_i}$

    `/* Update the original magnitude */`

    if $(D(G_{sp_i})$ *is in sector* $k)$ **then**

      $originalMag = (1 - \lambda - \tau) \times originalMag + \lambda \times avgNeighborMag + \tau \times M(G_{sp_i})$

    else

      $originalMag = (1 - \lambda) \times originalMag + \lambda \times avgNeighborMag$

    end

  end

end

we predict the future magnitude value of the vector (Figure 6.8 (C)). Finally, we repeat this process for every single sub-space and generate the future multi-vector field VF_{t+1} (Figure 6.8 (D)).

6.1.5 Missing-Data Imputation

We propose another approach to reduce the data sparsity issue so that we improves the performance of our prediction model. Since the movement data is sparse, the time series which is underlying data for forecasting described in Section 6.1.4 would be not continuous. The missing-data has a significant effect on the forecasting results, since it can increase the impact of data uncertainty in the forecasting process. For example, the graph in Figure 6.9 (Top) shows the magnitude values of a specific direction in a cell for 150 days with 4 hours time window. There are many missing data points that can cause inaccurate data forecasting. Thus, we apply a data imputation technique to reduce the data sparsity impact. We utilize spline interpolation technique [140] which is one of popular polynomial interpolation techniques to replace the missing data points with estimated value based on other available data points. The graph in Figure 6.9 (Bottom) shows the interpolation result. The blue circles are the observed data points and the red ones are imputed ones. We use the interpolated magnitude values to forecast the future value. In result, this data imputation approach improves the forecasting accuracy. More detailed evaluation results are explained in Section 6.2.

6.1.6 Visualization of Multi-Vector Fields

The flow of human crowds represents the temporal trend of human movement within a certain time period. Our system is built on several vector field visualization techniques for the generated multi-vector field data in order to observe the trends and patterns. In this work, we utilize the particle advection technique [141] for the vector fields and extend the web based project, *earth*, that visualizes global weather conditions [142]. Our web-based flow visualization system consists of JavaScript and several APIs, such as *D3.js*, *Back-*

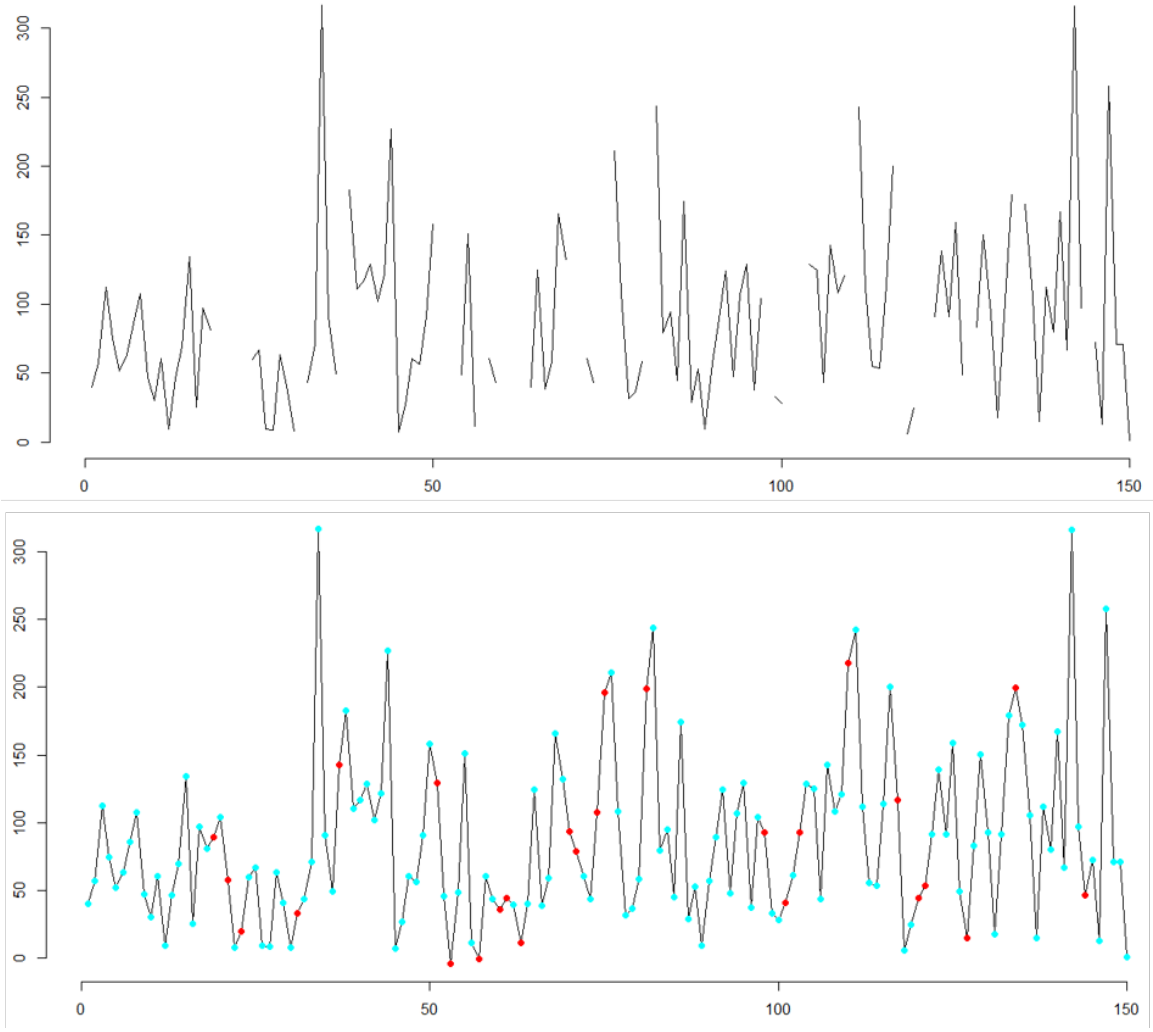

Fig. 6.9. Magnitude values of a specific direction in a grid cell for 150 days with 4 hours (Top). Interpolated result (Bottom). The blue circles are the observed data points and the red ones are imputed ones.

bone.js, and *When.js*. The web server visualizes the 3D globe according to the vector fields using D3 projections. The server, then, attaches some minimum geographical information including roads, country boundaries, and lakes using *TopoJSON*. The web-based flow visualization provides animated particles and the color of a particle varies accordingly as the particle ages. If the target vector field area is too small to be visible, our system allows

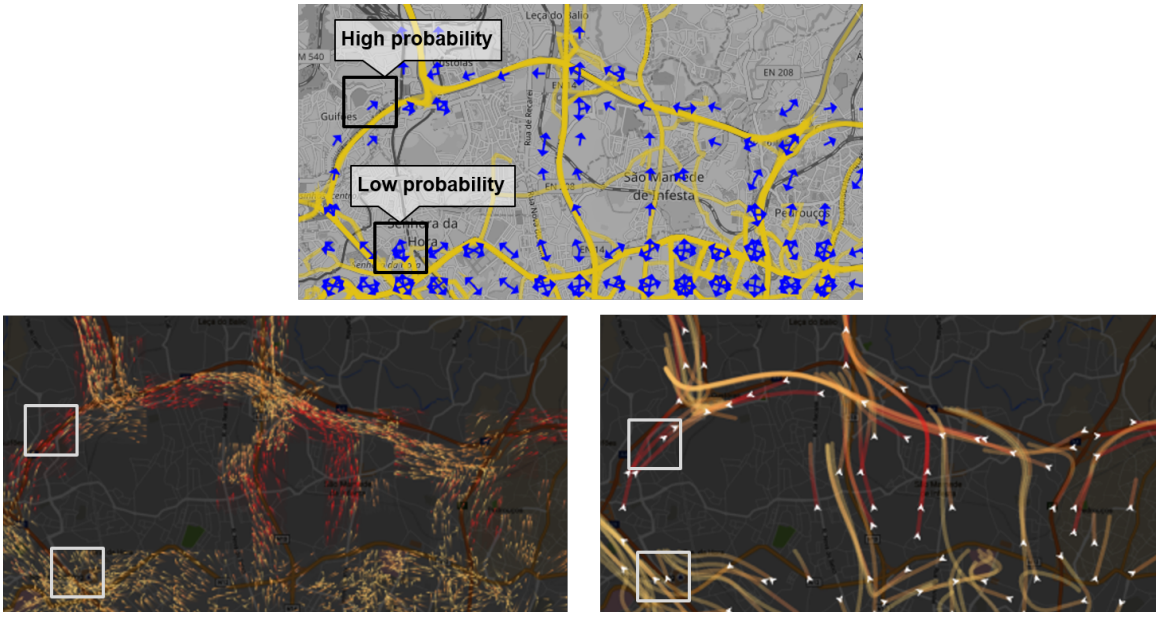

Fig. 6.10. The flows of Taxi data between 7:00 AM and 9:00 AM. Multi-vector fields representing the directional densities (Top). Particle advection (Left-Bottom). The paths of long-life particles (Right-Bottom). The red color indicates a high probability, whereas, the yellow represents a low probability.

users to apply an additional map layer located between the globe layer and the particle layer.

In Section 6.1.2, we introduce our multi-vector fields by the directional density estimation preserving the original vector directions instead of one dominant or average direction. Each direction has a density giving a hint for the probability of moving toward the direction. Since each grid cell can have multiple directions in this work, the densities for the directions can be normalized and used as a probability. Figure 6.10 (Top) shows an example of the multi-vector fields representing the directional densities for taxi flows in Porto, Portugal during 2 hours from 7:00 AM to 9:00 AM. In Figure 6.10 (Top), we highlight a region with a high probability direction (100%) and another region with 5 low probability directions (40%, 30%, 10%, 10%, 10%). A direction with high probability indicates that a flow is moving toward the direction with certainty whereas, a flow moving to a direction

with low probability visualized relatively uncertain flow. For each grid cell, we visualize the multi-vectors representing the directions of the moving flows based on its probability distribution.

In order to represent the multi-vector fields in our flow visualization, particles are generated and then animated through the vector fields. The particle color is determined by the probability of the movement direction of the particle, where the color gradually varies from red (high probability) to yellow (low probabiloty). An example of probability flow is shown in Figure 6.10 (Left-Bottom). The probability flow is created based on the probabilities in the multi-vector field. Particles are randomly generated in the space, and they move toward all non-zero directions in the multi-vector field. The number of particles in a given direction is proportional to the probability. For example, if there are two non-zero vectors (90% and 10%) in the multi-vector field and 100 particles are passing through the grid cell, 90 particles move in the 90% vector direction, whereas, ten particles move in 10% vector direction.

Also, when a particle enters a new cell, the probability of the particle path is multiplied by the prior probability. In this way, we can compute all probabilities along the particle path from its birth to death. The probability of each path segment is encoded in the path color as mentioned above. Since the particle path is obtained by using multi-vector fields, the probability flow tends to become complex as combinations of all the flow directions between grid cells are visualized at the same time. To reduce this limitation, we provide another type of visualization. Base on the lifespan of each particle, we connect the paths of the long-life particles. We can see more continuous and obvious paths over the space as shown in Figure 6.10 (Right-Bottom). For this visualization, we use the same color scheme as the previous one. The color represents the certainty/uncertainty of the flow based on the probability of the path segment.

6.2 Evaluation

We evaluate the performance of our forecasting model. For this experiment, we use taxi trajectory data which includes the trajectories for all the 442 taxis running in the city of Porto, in Portugal [143] and geo-location Twitter data generated around the New York City. We use the Normalized Root Mean Square Error (NRMSE) [144] to measure the error rate, instead of the Root Mean Square Error (RMSE) to help mitigate for the influence of outliers that generate a low reliability in the evaluation and for scale-independent. The measure is also widely applicable, and easily interpretable.

To compute NRMSE, we first define two time series: (1) Historical time series defined by $Y_k = \{y_t : t \in T\}$, where $k = \{1, \cdots, S\}$, and (2) Forecasted time series given by $\hat{Y}_k = \{\hat{y}_t : t \in T\}$. NRMSE can be calculated as follows:

$$NRMSE = \frac{1}{y_{max} - y_{min}} \sqrt{\frac{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}{T}} \tag{6.2}$$

Approach Comparison: We investigate the effect of the different approaches on the forecasting results. We compare the forecasting results by between the approach 1 including the fixed direction classification (Section 6.1.3) and the smoothing technique (Section 6.1.1) and the approach 2 including the method considering road directions (Section 6.1.2) and the data imputation technique (Section 6.1.5). We evaluate the impact on the forecasting accuracy in different grid size conditions. Figure 6.11 shows the evaluation results. The light blue bars represent the results of the approach 1, and orange bars represent the results by the approach 2. As shown in the chart, the forecasting accuracy is significantly improved under every grid size conditions when we use the approach 2, comparing the results of the approach 1. The approach 2 reduces the data uncertainty by using more realistic directions of the moving objects and imputing the missing data points.

Varying Grid Sizes: We investigate the error rates on different granularity level, different grid sizes (e.g., 200m, 500m, 1km, 2km, and 4km) as shown in Figures 6.12. We use the approach 2 for every grid size condition. The light blue bars represent the error rates for the taxi data and orange bars represent the results for the Twitter data. The error

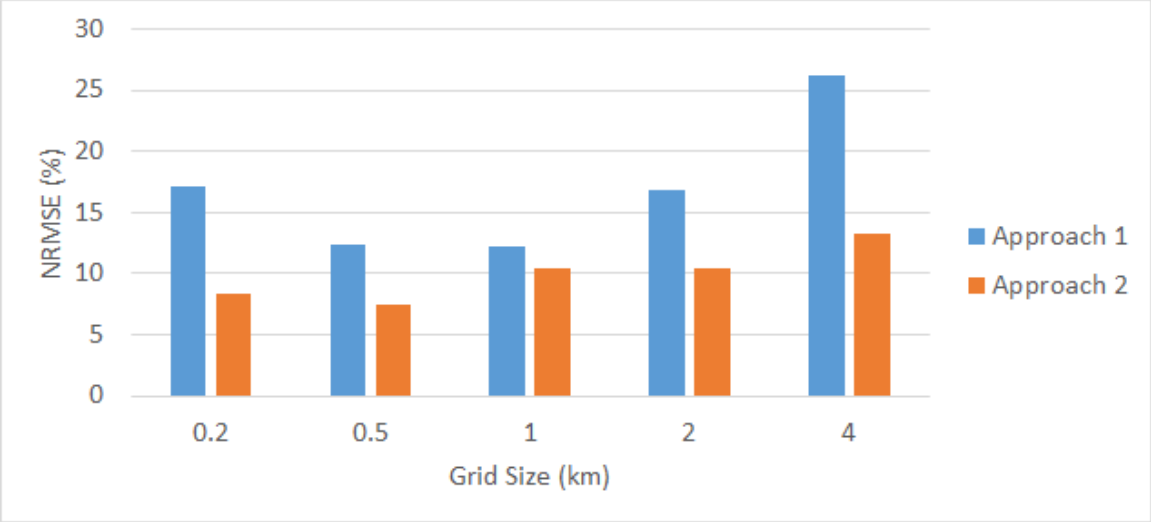

Fig. 6.11. Comparison of forecasting results by two different approaches.

rates of Twitter data are higher than ones of the taxi data for every condition. As the taxi data is more reliable and significantly low sparsity than the twitter data, it decreases its uncertainty. Also, we can see when we use the grid size as 0.5 km, we can obtain the best results for the both datasets —4.9% for the taxi data and 8.4% for the Twitter. We find out both error rates increase as the grid size grows. It shows that a larger area can have higher uncertainty regarding movement direction. However, the results of 0.2 km is worse than 0.5 km, that is, we model more fine grain data chunks when we use 0.2 km as the grid size, so the impact of data sparsity increases. We discuss these results further in Section 6.3.

Method Impact Comparison: The approach 2 includes the method considering road directions and the missing-data imputation method. We investigate the degree of the impact of the two different methods in order to improve the methods and find new methods increasing the forecasting accuracy. The chart in Figure 6.13 shows the error rates in varying grid sizes for the Twitter data under the different combinations of the methods. In the chart, R is the method considering road conditions, I is the missing-data imputation method, and R+I is the combination of the both methods. For every condition, we can the both methods improves the forecasting accuracy, but the use of R method has a higher impact on the

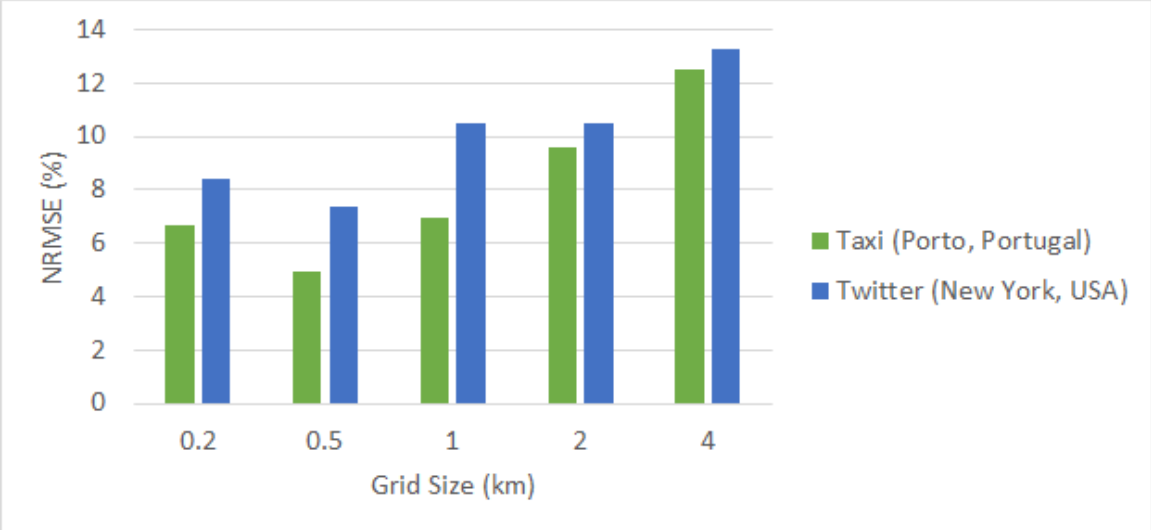

Fig. 6.12. Error rates for varying grid sizes for the Twitter and the taxi datasets.

forecasting accuracy than the use of I method because when we use the R method only, the error rates are lower than when we use the I method only. The use of R+I method provides the lowest error rate for every grid size condition. We also conduct a similar experiment on the taxi data. We measure only the impact of the I method since the taxi data is a set of high-resolution trajectories with geo-location points. Figure 6.14 shows the result for the taxi data. In this case, the use of R+I method also provides the better result than the non-use of I method. Based on these experimental results, we believe that when the data has less uncertainty, which means that it reflects actual or more realistic mobility, the performance of our forecasting model improves.

Visual Comparison: We also visually compare the visualization results created by the advanced visualization techniques described in Section 6.1.6 between observed data and forecasting result. Figure 6.15 shows the comparison result (Top: Observed data, Bottom: Forecasting data) by the particle advection with arrows in Porto. For this case, we use 0.5 km as the grid size; the NRMSE is 4.99%. As Figure 6.15 is a snapshot captured from animated visualization, it is not easy to compare this type of visualization. We can effectively compare the two results of the animated version. The visualization result of the forecasting

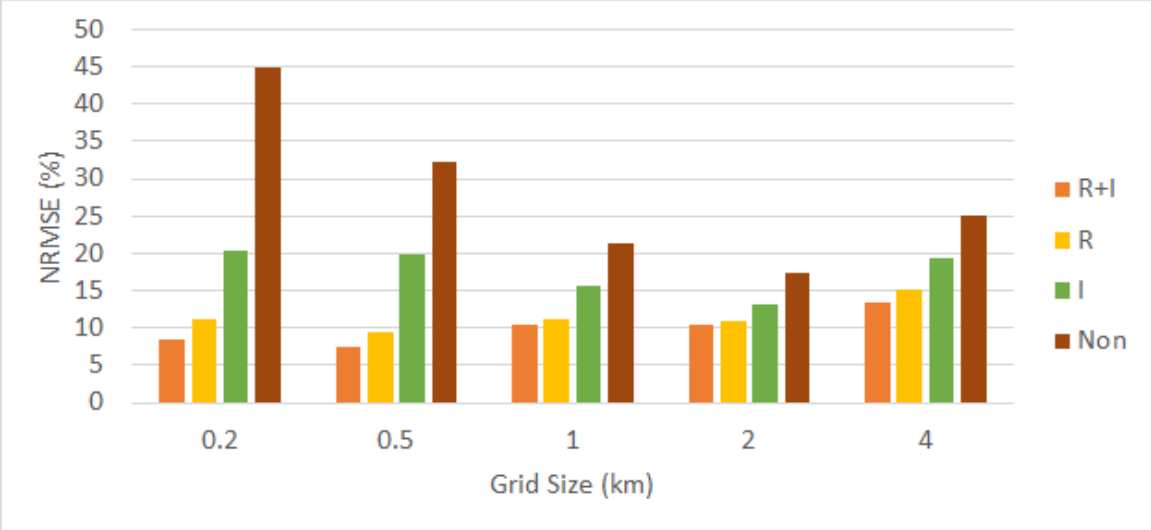

Fig. 6.13. Error rates for varying grid sizes for the Twitter data under the different combinations of the methods. R is the method considering road conditions and I is the missing-data imputation method.

data shows significantly similar patterns to the actual result patterns. Figure 6.16 shows the comparison result (Top: Observed data, Bottom: Forecasting data) by the paths of long-life particles in Porto. For this visualization result, we can see some different patterns around the bottom area. The bottom area has a higher density regarding movement and direction than the upper area as shown in Figure 6.10 (Top). When generating the paths of the long-life particles, our algorithm considers the more number of particles in the high-density areas. It would make different paths between the two visualization results. Figure 6.17 shows the comparison result (Left: Observed data, Right: Forecasting data) by the particle advection in New York City. For this case, we use 0.5 km as the grid size; the NRMSE is 7.4%. Figure 6.18 shows the comparison result (Left: Observed data, Right: Forecasting data) by the paths of long-life particles in New York City. We can see flows move from the north to south in Manhattan and enter inside Manhattan in both results. However, the flows in the middle of Manhattan are different between the observed and forecasting visualization results because of the same reason of the taxi data in Porto.

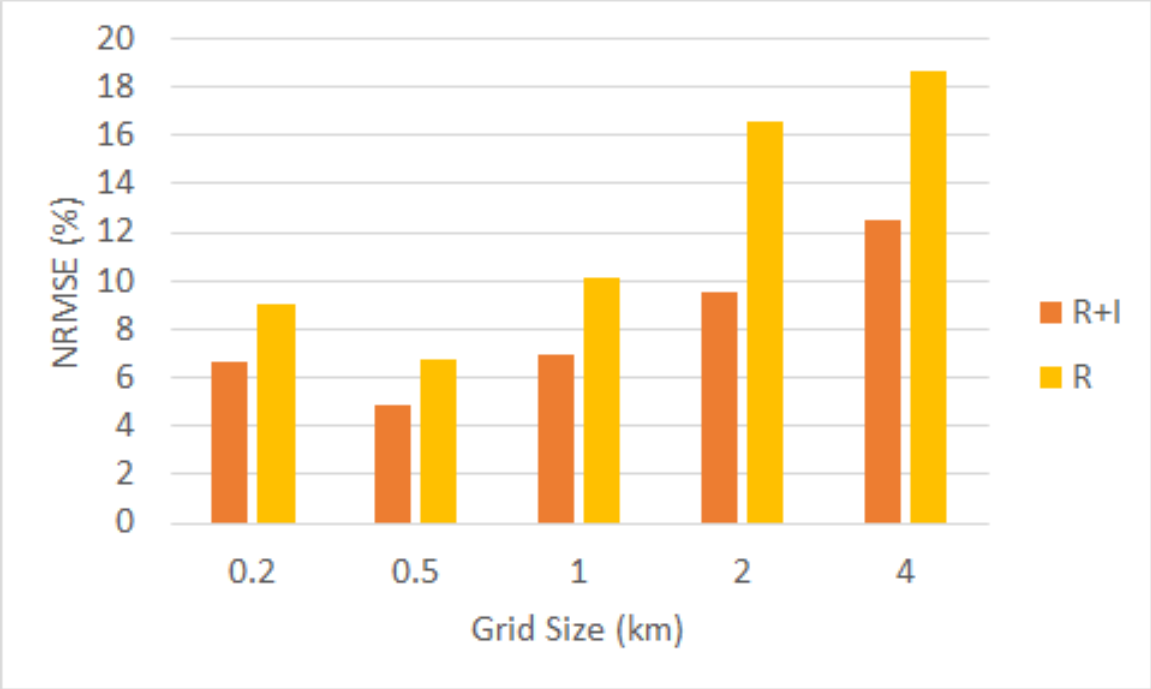

Fig. 6.14. Error rates for varying grid sizes for the taxi data under the different approach conditions.

6.2.1 Spatial Error Analysis

We evaluate which locations have high error and low error so that we can investigate whether the underlying characteristics of the location have a strong impact on the forecasting accuracy (e.g., road network characteristics (e.g., direction, speed limit), traffic densities, and the land-use (e.g., residential or industrial area)). In order to assess the error rates, we compute NRMSE for each location separately, and the values are normalized by the range of the NRMSE values. The normalized values are mapped to different colors ranging from green (low error) to red (high error). Figure 6.19 shows an example result of the spatial analysis on taxi trajectories in the city of Porto, Portugal. We calculate the error rate of each grid cell and fill the grid with a specific color between green and red according to its error rate as shown in Figure 6.19 (Left). The visualization in Figure 6.19 (Left) clearly shows the global overview of the error rates of the space. We find that the locations

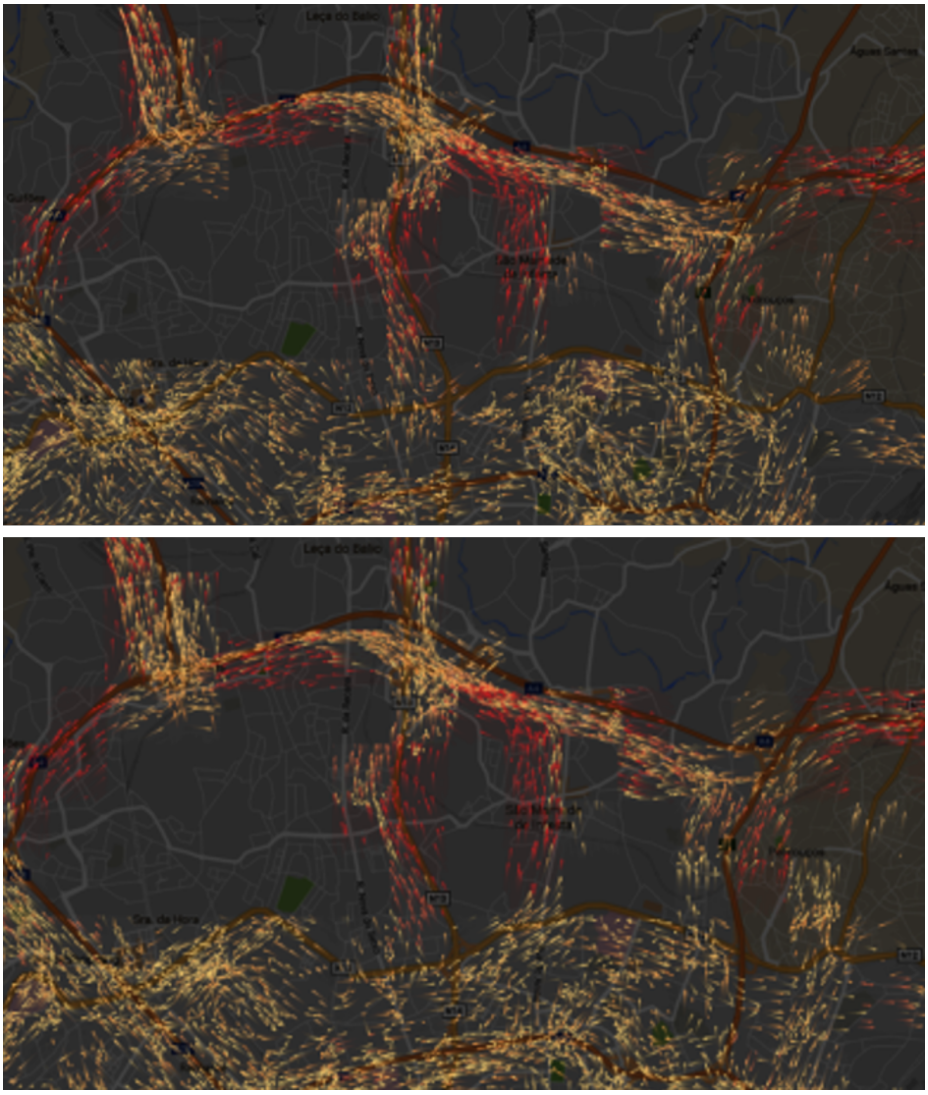

Fig. 6.15. Visual comparison of particle advection: Taxi in Porto. (Top: Observed data, Bottom: Forecasting data)

that are near the center of the city, where have many intersecting roads, have relatively high error rates. However, it is not easy to see the detailed geographical characteristics of each location, even though the colors are semi-transparent. To address this limitation, we provide another type of visualization. We draw a small circle at the top-left corner of each grid cell instead of filling the entire area using the same corresponding color. This visualization enables seeing the geographical features of the locations as well as the region's error rate.

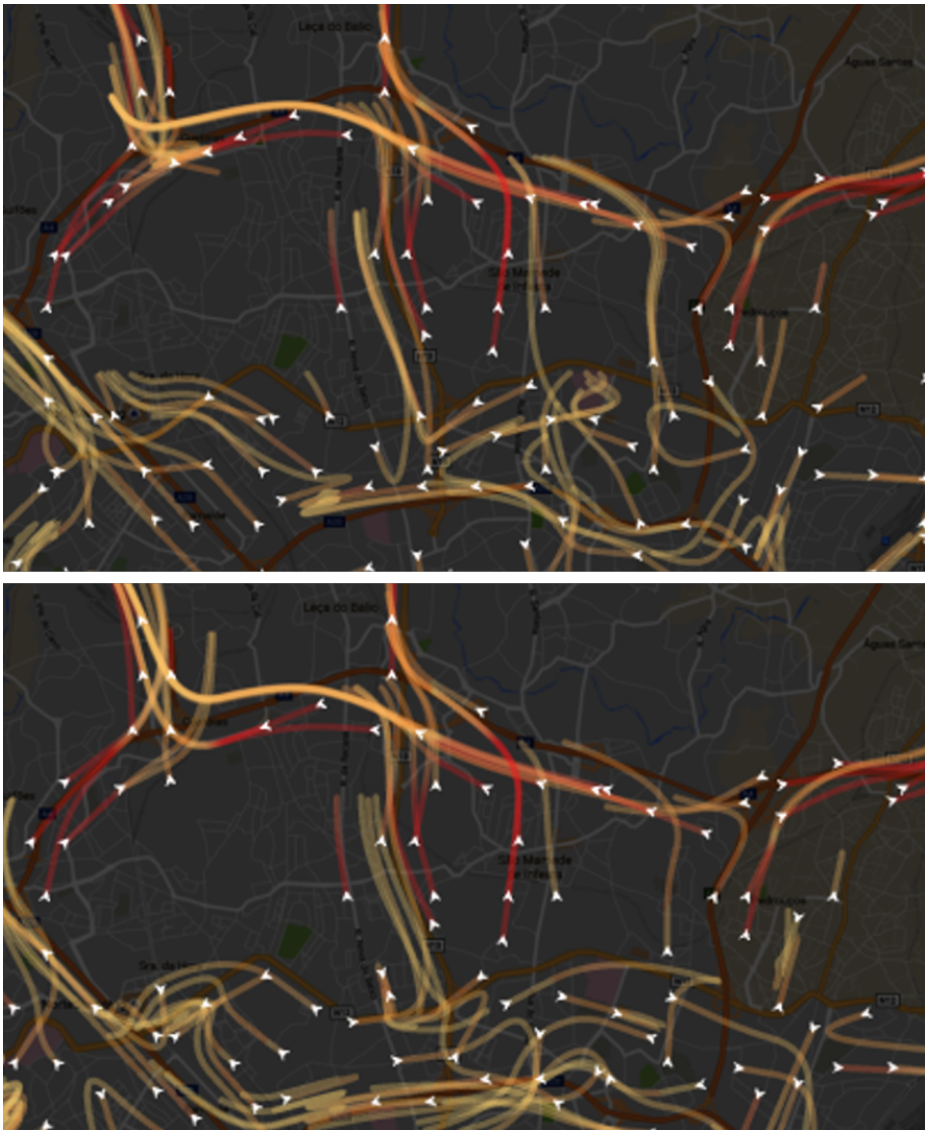

Fig. 6.16. Visual comparison of paths of long-life particles: Taxi in Porto.
(Top: Observed data, Bottom: Forecasting data)

The users can switch between the two types of visualizations. We discuss these results further in Section 6.3.2.

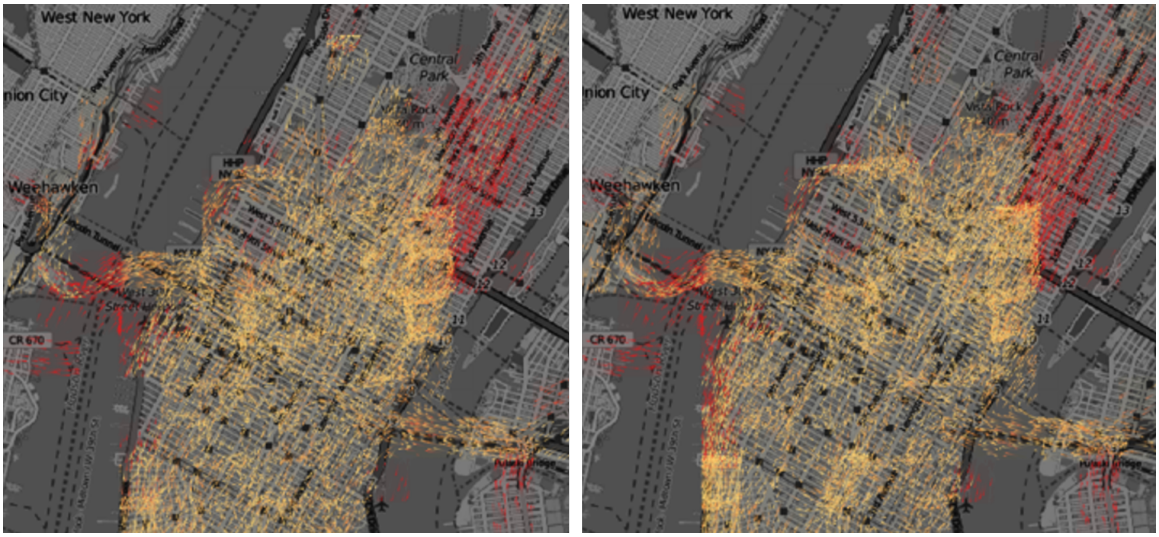

Fig. 6.17. Visual comparison of particle advection: Twitter in New York City. (Left: Observed data, Right: Forecasting data)

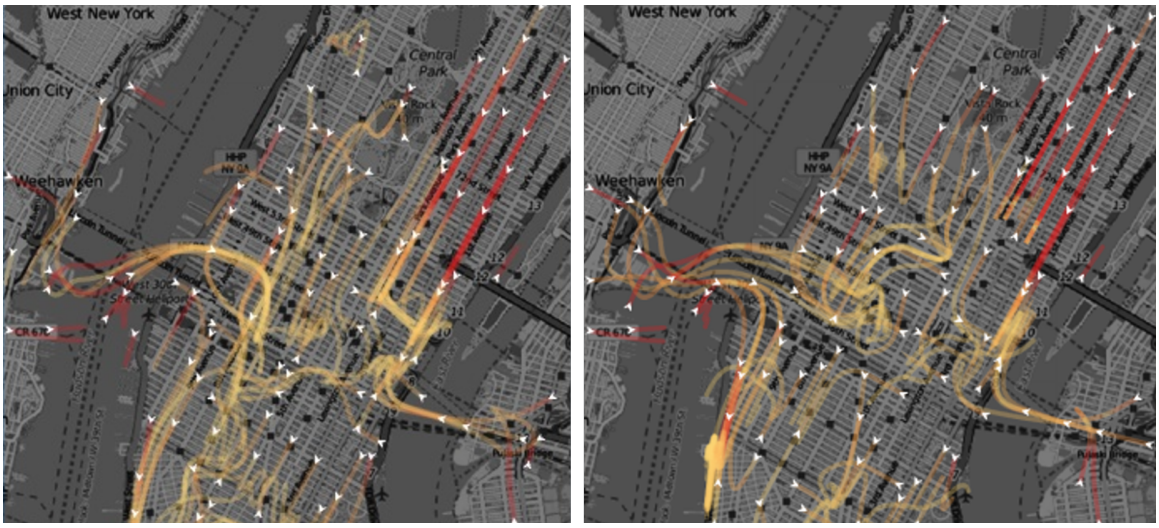

Fig. 6.18. Visual comparison of paths of long-life particles: Twitter in New York City. (Left: Observed data, Right: Forecasting data)

6.3 Discussion

In this section, we discuss issues related to the grid size and regional error rate differentials for our methodology. Specifically, we discuss the challenges associated with selecting the appropriate grid size and regional influence on the forecasting accuracy.

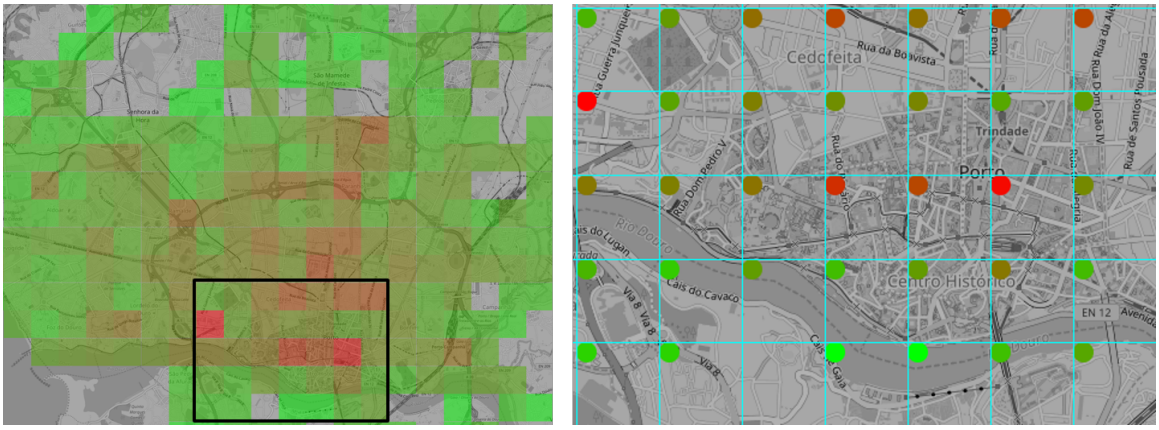

Fig. 6.19. Spatial error analysis: Regional differentials in error rate. Red color indicates the high error rate; Green color indicates the low error rate.

6.3.1 Grid Size

As discussed in Section 6.1.2, we conducted our evaluations by varying the grid size. Our evaluation results reveal that the grid size has an impact on the accuracy of the forecast results. However, selecting an appropriate grid size (i.e., geospatial scale) for analysis remains to be a challenging task. Based on our experiment results, in general, the data with high density is less vulnerable and the accuracy of forecasting with a higher granularity (fine scale) is higher than that with low granularity (coarse scale). One adverse case happens when the grid size is 200 meters where we find that the forecast accuracy is worse than that with 500 meters grid size as shown in Figure 6.12. Interestingly, this is observed in both Twitter and Taxi data. We find out that when we use 0.2 km as the grid size, the data sparsity increases, and the forecasting accuracy decreases. Accordingly, we believe that there may exit a relationship between the grid size and the geographical characteristics (e.g., demographics) that needs to be further explored. We leave this as future work.

6.3.2 Regional Differentials in Forecasting Error Rate

The spatial error analysis described in Section 6.2 helps in understanding the relationship between the forecasting accuracy and the regional characteristics. We find that the locations with roads that have more than 4 different directions and a higher traffic density may have higher error rates. On the other hand, we can observe that the locations with fewer roads choices and interactions have lower error rates, even though they have a higher traffic density. We believe that route diversity has a high impact on regular movement patterns, which affects forecasting the patterns. One of the possible ways to reduce this issue would be to adapt the grid size based on the route diversity of each location, instead of using a fixed grid size for every location. We need further investigation to understand how route diversity influences the forecasting accuracy and find possible solutions. We leave this as future work.

6.4 Summary

We presented a space-based visual analytics approach for forecasting the overall flow of human crowds. Our work utilizes individual movement trajectory data and embeds it into a two-dimensional Euclidean space. Our approach is based on modeling for the space instead of the more conventional approach of modeling individual objects. We propose a new method to estimate the directional density for representing the overall flows of moving objects. We also introduce a new model for forecasting the future flow of human crowds using the seasonal trend decomposition based on Loess technique. Finally, we develop a new flow visualization technique of multi-vector fields to represent the directional flow density.

7. CONCLUSIONS

In summary, we presented design and development of visual analytics techniques and systems for spatial decision support through coupling modeling of spatiotemporal social media data, with scalable and interactive visual environments. We extract valuable hidden information from the huge volume of unstructured social media data and model the extracted information for visualizing meaningful information along with user-centered interactive interfaces. In conclusion, we summarize the major contributions of this thesis as the following:

- **Visual analytics of location-based social networks for abnormal event detection:** We presented a visual analytics approach that provides users with scalable and interactive social media data analysis and visualization including the exploration and examination of abnormal topics and events within various social media data sources, such as Twitter, Flickr and YouTube. We also introduced an interactive visual spatial decision support environment that assists in evacuation planning and disaster management.

- **Visual analytics for public behavior analysis in disaster events:** We demonstrated how to improve investigation by analyzing the extracted public behavior responses from social media before, during and after natural disasters, such as hurricanes and tornadoes.

- **Visual analytics of anomalous human movement analysis:** We proposed a trajectory-based visual analytics system for analyzing anomalous human movements during disasters using multi-online media. We discover common human movement patterns using extracted trajectories from LBSNs and propose a classification model for detecting abnormal movements. In addition, we integrate multiple visual repre-

sentations using relevant context extracted from different online media sources for improving situational awareness.

- **Visual analytics of forecasting the flow of human crowds:** We introduced a novel space-based visual analytics approach for forecasting the overall flow of these human crowds. We apply seasonal trend analysis techniques on the flow data in each subspace to forecast the future level of crowd movement based on the observed historical time series flow patterns. Our methodology is comprised of directional flow density estimation techniques for preserving original paths and movement directions, and a novel flow smoothing method utilizing local and global trends to mitigate data sparsity and noise issues.

7.1 Future Work

Although we have shown effectiveness of our visual analytics techniques and systems, there are some other aspects we have not treated in detail. Furthermore, as data becomes complex, new challenges arise. Given the visual analytics techniques for location-based social networks that my past research has generated, my future research directions consist of the following topics:

- **Advanced anomaly Detection:** We have presented a visual analytics approach that provides users with visual analytics techniques including the exploration and examination of abnormal topics and events within various social media data sources. We need to further investigate context-based analysis and improve the current detection algorithm to allow for a faster analysis. Due to the fast-paced and low quality nature of micro-blogging, the effects of additional pre-processing options like automated spell-checking or synonym recognition under the constraint of preventing ambiguities should be considered. Further research is also required to supplement the system with real-time monitoring features, demanding additional means for adaptive attention guiding as well as interaction techniques for use in high pressure environments.

- **Enhanced visualization:** We have shown a trajectory-based visual analytics system for anomalous human movement analysis during disasters using multi-type online media. We have limitations in reducing the visual clutter of trajectories and adding annotations on the map. The classification model to automatically identify the abnormal movements is required. For the forecasting flow visualization, our framework has used fixed directions for each grid. However, wee need to incorporate data-driven methods to consider road directions.

- **Improved analysis for data sparsity:** Sparse and noisy flow data often generate non-smooth flow patterns that cannot be used for accurate prediction. In order to mitigate the effect caused by the data sparseness and noise, we proposed a new flow smoothing method based on local and global trend estimation. However, this smoothing technique has limitation to generate aspects, because it does not consider the road network. We plan to investigate other approach to reduce the data sparsity issue. In addition, further research is need that investigate the effects of data sparsity and noise issues on our forecasting results.

- **Evaluation of visual analytics tools within the crisis management:** Finally, we plan conduct a user evaluation for the usability and effectiveness of the geospatial visual support, and the impact of interactive spatiotemporal visual analytics using social media data with crisis management personnel and other domain experts.

LIST OF REFERENCES

LIST OF REFERENCES

[1] J. J. Thomas and K. A. Cook, eds., *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[3] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "Stl: A seasonal-trend decomposition procedure based on loess (with discussion)," *Journal of Official Statistics*, vol. 6, pp. 3–73, 1990.

[4] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pp. 336 –343, sep 1996.

[5] Committee on Public Response to Alerts and Warnings Using Social Media: Current Knowledge and Research Gaps; Computer on Science and Technology Board; Division on Engineering and Physical Sciences; National Research Council, *Public Response to Alerts and Warnings Using Social Media: Report of a Workshop on Current Knowledge and Research Gaps*. The National Academies Press, 2013.

[6] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proceedings of the 13th international conference on Ubiquitous computing*, pp. 89–98, ACM, 2011.

[7] L.-Y. Wei, Y. Zheng, and W.-C. Peng, "Constructing popular routes from uncertain trajectories," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 195–203, ACM, 2012.

[8] S. Eubank, H. Guclu, V. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, "Modelling disease outbreaks in realistic urban social networks," *Nature*, vol. 429, no. 6988, pp. 180–184, 2004.

[9] G. Andrienko, N. Andrienko, P. Bak, S. Kisilevich, and D. Keim, "Analysis of community-contributed space-and time-referenced data (example of panoramio photos)," in *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pp. 540–541, ACM, 2009.

[10] G. Fuchs, G. Andrienko, N. Andrienko, and P. Jankowski, "Extracting personal behavioral patterns from geo-referenced tweets," in *AGILE*, 2013.

[11] L. Gabrielli, S. Rinzivillo, F. Ronzano, and D. Villatoro, "From tweets to semantic trajectories: mining anomalous urban mobility patterns," in *Citizen in Sensor Networks*, pp. 26–35, Springer, 2014.

[12] N. Hochman and R. Schwartz, "Visualizing instagram: Tracing cultural visual rhythms," in *Proceedings of the Workshop on Social Media Visualization (SocMedVis) in conjunction with the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM–12)*, pp. 6–9, 2012.

[13] T. T. Zin, P. Tin, H. Hama, and T. Toriu, "Knowledge based social network applications to disaster event analysis," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2013.

[14] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel, "From movement tracks through events to places: Extracting and characterizing significant places from mobility data," in *IEEE Symposium on Visual Analytics Science and Technology*, pp. 161–170, IEEE, 2011.

[15] T. Fujisaka, R. Lee, and K. Sumiya, "Discovery of user behavior patterns from geotagged micro-blogs," in *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, p. 36, ACM, 2010.

[16] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *ACM SIGMOD international conference on Management of data*, pp. 593–604, ACM, 2007.

[17] B. D. Dalziel, B. Pourbohloul, and S. P. Ellner, "Human mobility patterns predict divergent epidemic dynamics among cities," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 280, no. 1766, p. 20130763, 2013.

[18] J. Bao, Y. Zheng, and M. F. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pp. 199–208, ACM, 2012.

[19] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[20] A. Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu, "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction," in *IEEE International Conference on Data Engineering*, pp. 254–265, IEEE, 2013.

[21] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Semantic trajectory mining for location prediction," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 34–43, ACM, 2011.

[22] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[23] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "Wherenext: a location predictor on trajectory pattern mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 637–646, ACM, 2009.

[24] W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden markov models," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 911–918, ACM, 2012.

[25] A. Monreale, G. L. Andrienko, N. V. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel, "Movement data anonymity through generalization.," *Transactions on Data Privacy*, vol. 3, no. 2, pp. 91–121, 2010.

[26] L.-Y. Wei, Y. Zheng, and W.-C. Peng, "Constructing popular routes from uncertain trajectories," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 195–203, ACM, 2012.

[27] Z. Wang, T. Ye, M. Lu, X. Yuan, H. Qu, J. Yuan, and Q. Wu, "Visual exploration of sparse traffic trajectory data," *IEEE Trans. on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1813–1822, 2014.

[28] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel, *Visual analytics of movement*. Springer Publishing Company, Incorporated, 2013.

[29] R. L. Hughes, "The flow of human crowds," *Annual review of fluid mechanics*, vol. 35, no. 1, pp. 169–182, 2003.

[30] A. MacEachren, A. Jaiswal, A. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "Senseplace2: Geotwitter analytics support for situational awareness," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 181 –190, oct. 2011.

[31] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, WWW '10, (New York, NY, USA), pp. 851–860, ACM, 2010.

[32] N. Andrienko, G. Andrienko, H. Stange, T. Liebig, and D. Hecker, "Visual analytics for understanding spatial situations from episodic movement data," *KI-Künstliche Intelligenz*, vol. 26, no. 3, pp. 241–251, 2012.

[33] G. Andrienko, N. Andrienko, and S. Wrobel, "Visual analytics tools for analysis of movement data," *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 2, pp. 38–46, 2007.

[34] N. Andrienko and G. Andrienko, "A visual analytics framework for spatio-temporal analysis and modelling," *Data Mining and Knowledge Discovery*, vol. 27, no. 1, pp. 55–83, 2013.

[35] N. Andrienko and G. Andrienko, "Visual analytics of movement: An overview of methods, tools and procedures," *Information Visualization*, pp. 3–24, 2013.

[36] G. L. Andrienko, N. V. Andrienko, J. Dykes, S. I. Fabrikant, and M. Wachowicz, "Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research," *Information Visualization*, vol. 7, no. 3-4, pp. 173–180, 2008.

[37] P. Lundblad, O. Eurenius, and T. Heldring, "Interactive visualization of weather and ship data," in *13th International Conference on Information Visualisation*, pp. 379–386, IEEE Computer Society, 2009.

[38] C. Tominski, H. Schumann, G. L. Andrienko, and N. V. Andrienko, "Stacking-based visualization of trajectory attribute data," *IEEE Trans. Vis. Comput. Graph*, vol. 18, no. 12, pp. 2565–2574, 2012.

[39] W. R. Tobler, "Experiments in migration mapping by computer," *The American Cartographer*, vol. 14, no. 2, pp. 155–163, 1987.

[40] T. Kapler and W. Wright, "Geotime information visualization," *Information Visualization*, vol. 4, no. 2, pp. 136–146, 2005.

[41] C. Hurter, B. Tissoires, and S. Conversy, "Fromdady: Spreading aircraft trajectories across views to support iterative queries," *IEEE Trans. on Visualization and Computer Graphics*, vol. 15, pp. 1017–1024, Nov. 2009.

[42] N. Ferreira, J. T. Klosowski, C. E. Scheidegger, and C. T. Silva, "Vector fields: Vector field *k*-means: Clustering trajectories by fitting multiple vector fields," *Computer Graphics Forum*, vol. 32, pp. 201–210, June 2013.

[43] G. L. Andrienko, N. V. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti, "Interactive visual clustering of large collections of trajectories," in *IEEE VAST*, pp. 3–10, IEEE Computer Society, 2009.

[44] N. Andrienko, G. Andrienko, L. Barrett, M. Dostie, and P. Henzi, "Space transformation for understanding group movement," *IEEE Trans. Visualization and Computer Graphics*, vol. 19, pp. 2169–2178, Dec 2013.

[45] R. Scheepens, H. van de Wetering, and J. J. van Wijk, "Non-overlapping aggregated multivariate glyphs for moving objects," in *IEEE Pacific Visualization Symposium, PacificVis 2014, Yokohama, Japan, March 4-7, 2014*, pp. 17–24, IEEE, 2014.

[46] N. Willems, H. van de Wetering, and J. J. van Wijk, "Visualization of vessel movements," *Comput. Graph. Forum*, vol. 28, no. 3, pp. 959–966, 2009.

[47] J. Wood, J. Dykes, and A. Slingsby, "Visualisation of origins, destinations and flows with OD maps," *The Cartographic Journal*, vol. 47, no. 2, pp. 117–129, 2010.

[48] D. Guo and X. Zhu, "Origin-destination flow data smoothing and mapping," *IEEE Trans. Vis. Comput. Graph*, vol. 20, no. 12, pp. 2043–2052, 2014.

[49] I. Boyandin, E. Bertini, P. Bak, and D. Lalanne, "Flowstrates: An approach for visual exploration of temporal origin-destination data," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 971–980, 2011.

[50] J. Poco, H. Doraiswamy, H. T. Vo, J. a. L. D. Comba, J. Freire, and C. T. Silva, "Exploring traffic dynamics in urban environments using vector-valued functions," in *Proceedings of the 2015 Eurographics Conference on Visualization*, EuroVis '15, (Aire-la-Ville, Switzerland, Switzerland), pp. 161–170, Eurographics Association, 2015.

[51] J. C. Nascimento, M. A. Figueiredo, and J. S. Marques, "Trajectory analysis in natural images using mixtures of vector fields," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 4353–4356, IEEE, 2009.

[52] D. Chu, D. A. Sheets, Y. Zhao, Y. Wu, J. Yang, M. Zheng, and G. Chen, "Visualizing hidden themes of taxi movement with semantic transformation," in *IEEE Pacific Visualization Symposium, PacificVis 2014, Yokohama, Japan, March 4-7, 2014*, pp. 137–144, IEEE, 2014.

[53] G. L. Andrienko, N. V. Andrienko, G. Fuchs, A.-M. O. Raimond, J. Symanzik, and C. Ziemlicki, "Extracting semantics of individual places from movement data by analyzing temporal patterns of visits," in *ACM SIGSPATIAL International Workshop on Computational Models of Place*, pp. 9–15, 2013.

[54] R. Krueger, D. Thom, and T. Ertl, "Visual analysis of movement behavior using web data for context enrichment," in *Proceedings of the 2014 IEEE Pacific Visualization Symposium*, PACIFICVIS '14, pp. 193–200, IEEE Computer Society, 2014.

[55] W. Wu, J. Xu, H. Zeng, Y. Zheng, H. Qu, B. Ni, M. Yuan, and L. M. Ni, "Telcovis: Visual exploration of co-occurrence in urban human mobility based on telco data," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 935–944, 2016.

[56] Y. Zheng and X. Zhou, eds., *Computing with Spatial Trajectories*. Springer, 2011.

[57] M. Dörk, S. Carpendale, C. Collins, and C. Williamson, "VisGets: Coordinated visualizations for web-based information exploration and discovery," *IEEE Transactions on Visualization and Computer Graphics (Proceedings Information Visualization 2008)*, vol. 14, no. 6, pp. 1205–1212, 2008.

[58] K. Field and J. O'Brien, "Cartoblography: Experiments in using and organising the spatial context of micro-blogging," *Transactions in GIS*, vol. 14, pp. 5–23, 2010.

[59] E. Roth and J. White, "Twitterhitter: Geovisual analytics for harvesting insight from volunteered geographic information," in *Proceedings of GIScience*, 2010.

[60] S. Wakamiya, R. Lee, and K. Sumiya, "Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, (New York, NY, USA), pp. 77–84, ACM, 2011.

[61] H. Bosch, D. Thom, M. Worner, S. Koch, E. Puttmann, D. Jackle, and T. Ertl, "Scatterblogs: Geo-spatial document analysis," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 309 –310, 2011.

[62] Bikely. `http://www.bikely.com/`.

[63] EveryTrail. `http://www.everytrail.com/`.

[64] J. J.-C. Ying, W.-C. Lee, M. Ye, C.-Y. Chen, and V. S. Tseng, "User association analysis of locales on location based social networks," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, (New York, NY, USA), pp. 69–76, ACM, 2011.

[65] P. C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner, and J. Thomas, "IN-SPIRE Infovis 2004 contest entry," in *IEEE Symposium on Information Visualization*, Oct. 2004.

[66] J. Weng and B.-S. Lee, "Event detection in twitter," in *International AAAI Conference on Weblogs and Social Media*, 2011.

[67] R. Lee and K. Sumiya, "Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, LBSN '10, (New York, NY, USA), pp. 1–10, ACM, 2010.

[68] A. Pozdnoukhov and C. Kaiser, "Space-time dynamics of topics in streaming text," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, (New York, NY, USA), pp. 1–8, ACM, 2011.

[69] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pp. 338–349, Berlin, Heidelberg: Springer-Verlag, 2011.

[70] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, (New York, NY, USA), pp. 261–270, ACM, 2010.

[71] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, (New York, NY, USA), pp. 306–315, ACM, 2004.

[72] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *ICWSM*, 2010.

[73] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, (Stroudsburg, PA, USA), pp. 248–256, Association for Computational Linguistics, 2009.

[74] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "Paralleltopics: A probabilistic approach to exploring document collections," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 231 –240, oct. 2011.

[75] D. Thom, H. Bosch, S. Koch, M. Woerner, and T. Ertl, "Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages," in *IEEE Pacific Visualization Symposium (PacificVis)*, 2012.

[76] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, "Twitinfo: aggregating and visualizing microblogs for event exploration," in *Proceedings of the 2011 annual conference on Human factors in computing systems*, pp. 227–236, ACM, 2011.

[77] N. Andrienko, G. Andrienko, and P. Gatalsky, "Exploring changes in census time series with interactive dynamic maps and graphics," *Computational statistics*, vol. 16, no. 3, pp. 417–433, 2001.

[78] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, "Semantics+ filtering+ search= twitcident. exploring information in social web streams," in *Proceedings of the 23rd ACM conference on Hypertext and social media*, pp. 285–294, ACM, 2012.

[79] T. Terpstra, A. de Vries, R. Stronkman, and G. Paradies, "Towards a realtime twitter analysis during crises for operational crisis management," in *Proc. of the 9th Inter. ISCRAM Conf*, 2012.

[80] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," in *Proceedings of the 28th international conference on Human factors in computing systems*, pp. 1079–1088, ACM, 2010.

[81] T. Heverin and L. Zach, "Microblogging for crisis communication: Examination of twitter use in response to a 2009 violent crisis in the seattle-tacoma, washington area," in *Proceedings of the 7th International ISCRAM Conference–Seattle*, vol. 1, 2010.

[82] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom, "Thematic patterns in georeferenced tweets through space-time visual analytics," *Computing in Science and Engg.*, vol. 15, pp. 72–82, May 2013.

[83] R. B. Braga, S. d. M. Medeiros da Costa, and H. Martin, "A trajectory correlation algorithm based on users' daily routines," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 501–504, ACM, 2011.

[84] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1082–1090, ACM, 2011.

[85] Z. Liao, Y. Yu, and B. Chen, "Anomaly detection in gps data based on visual analytics," in *IEEE Symposium on Visual Analytics Science and Technology*, pp. 51–58, IEEE, 2010.

[86] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The International Journal on Very Large Data Bases*, vol. 8, no. 3-4, pp. 237–253, 2000.

[87] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000.

[88] N. Adrienko and G. Adrienko, "Spatial generalization and aggregation of massive movement data," *IEEE Trans. Visualization and Computer Graphics*, vol. 17, pp. 205–219, Feb 2011.

[89] P. Laube, *Computational movement analysis*. Springer, 2014.

[90] N. J. Yuan, Y. Wang, F. Zhang, X. Xie, and G. Sun, "Reconstructing individual mobility from smart card transactions: A space alignment approach," in *IEEE International Conference on Data Mining*, pp. 877–886, IEEE, 2013.

[91] R. Assam and T. Seidl, "Check-in location prediction using wavelets and conditional random fields," in *IEEE International Conference on Data Mining*, pp. 713–718, IEEE, 2014.

[92] B. Kim, J.-Y. Ha, S. Lee, S. Kang, Y. Lee, Y. Rhee, L. Nachman, and J. Song, "Adnext: a visit-pattern-aware mobile advertising system for urban commercial complexes," in *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, pp. 7–12, ACM, 2011.

[93] N. Andrienko, G. Andrienko, and S. Rinzivillo, "Leveraging spatial abstraction in traffic analysis and forecasting with visual analytics," *Information Systems*, vol. 57, pp. 172–194, 2016.

[94] T. von Landesberger, F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren, "Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 11–20, 2016.

[95] M. Lu, Z. Wang, J. Liang, and X. Yuan, "Od-wheel: Visual design to explore od patterns of a central region," in *2015 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 87–91, IEEE, 2015.

[96] Twitter, "200 million tweets per day." Retrieved March 1, 2012, `http://blog.twitter.com/2011/06/200-million-tweets-per-day.html`, 2011.

[97] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. Ebert, and T. Ertl, "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in *Visual Analytics Science and Technology, 2012 IEEE Conference on*, pp. 143–152, Oct.

[98] A. K. McCallum, "Mallet: A machine learning for language toolkit." http://mallet.cs.umass.edu, 2002.

[99] R. Krovetz, "Viewing morphology as an inference process," in *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, (New York, NY, USA), pp. 191–202, ACM, 1993.

[100] T. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

[101] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[102] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert, "A visual analytics approach to understanding spatiotemporal hotspots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 205–220, Mar. 2010.

[103] New York Times, "Wall street protest begins with demonstrators blocked." Retrieved June 25, 2012, `http://cityroom.blogs.nytimes.com/2011/09/17/wall-street-protest-begins-with-demonstrators-blocked`, 2011.

[104] New York Times, "Police arresting protesters on brooklyn bridge." Retrieved June 25, 2012, `http://cityroom.blogs.nytimes.com/2011/10/01/police-arresting-protesters-on-brooklyn-bridge`, 2011.

[105] New York Times, "Major unions join occupy wall street protest." Retrieved June 25, 2012, `http://www.nytimes.com/2011/10/06/nyregion/major-unions-join-occupy-wall-street-protest.html`, 2011.

[106] New York Times, "Occupy wall street protests worldwide." Retrieved June 25, 2012, `http://www.nytimes.com/2011/10/16/world/occupy-wall-street-protests-worldwide.html`, 2011.

[107] United States Geological Survey (USGS), "Magnitude 5.8 - virginia." Retrieved March 30, 2012, http://earthquake.usgs.gov/earthquakes/recenteqsww/Quakes/se082311a.php, 2011.

[108] L. Indvik, "East coasters turn to twitter during virginia earthquake." Retrieved March 30, 2012, http://mashable.com/2011/08/23/virginia-earthquake/, 2011.

[109] Huffingtonpost, "Washington monument: Did earthquake-damaged icon sink or tilt?." Retrieved March 25, 2012, http://www.huffingtonpost.com/2012/03/14/washington-monument-did-e_n_1344422.html, 2012.

[110] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, (New York, NY, USA), pp. 759–768, ACM, 2010.

[111] J. Mahmud, J. Nichols, and C. Drews, "Where is this tweet from? Inferring home locations of twitter users," in *International AAAI Conference on Weblogs and Social Media*, 2012.

[112] P. C. Young, D. J. Pedregal, and W. Tych, "Dynamic harmonic regression," *Journal of Forecasting*, vol. 18, no. 6, pp. 369–394, 1999.

[113] G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.

[114] B. Jiang, S. Liang, J. Wang, and Z. Xiao, "Modeling modis lai time series using three statistical methods," *Remote Sensing of Environment*, vol. 114, no. 7, pp. 1432–1444, 2010.

[115] J. Chae, D. Thom, Y. Jang, S. Y. Kim, T. Ertl, and D. Ebert, "Visual analytics of microblog data for public behavior analysis in disaster events," in *EuroVis Workshop on Visual Analytics*, pp. 67–71, 2013.

[116] Twitter, "The geography of tweets." Retrieved August 24, 2013, https://blog.twitter.com/2013/geography-tweets-3, 2013.

[117] Wikipedia, "Hurricane sany." Retrieved April 30, 2013, http://en.wikipedia.org/wiki/Hurricane_Sandy, 2012.

[118] Wikipedia, "Effects of hurricane sandy in new jersey." Retrieved June 11, 2013, http://en.wikipedia.org/wiki/Effects_of_Hurricane_Sandy_in_New_Jersey, 2012.

[119] G. Andrienko, N. Andrienko, P. Jankowski, D. Keim, M. Kraak, A. MacEachren, and S. Wrobel, "Geovisual analytics for spatial decision support: Setting the research agenda," *International Journal of Geographical Information Science*, vol. 21, no. 8, pp. 839–857, 2007.

[120] A. Robinson, R. Roth, J. Blanford, S. Pezanowski, and A. MacEachren, "Developing map symbol standards through an iterative collaboration process," *Environment and Planning B: Planning and Design*, vol. 39, no. 6, pp. 1034–1048, 2012.

[121] Wikipedia, "2013 moore tornado." Retrieved June 12, 2013, http://en.wikipedia.org/wiki/2013_Moore_tornado, 2013.

[122] Twitchy, "Convoy of hope: Glenn beck arrives in okla. with two truckloads of food, water and diapers." Retrieved August 26, 2013, http://twitchy.com/2013/05/21/convoy-of-hope-glenn-beck-arrives-in-okla-with-two-truckloads-of-food-water-and-diapers/, 2013.

[123] R. Maciejewski, R. Hafen, S. Rudolph, S. Larew, M. Mitchell, W. Cleveland, and D. Ebert, "Forecasting hotspots - a predictive analytics approach," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 4, pp. 440–453, April.

[124] A. Malik, R. Maciejewski, E. Hodgess, and D. Ebert, "Describing temporal correlation spatially in a visual analytics environment," in *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pp. 1–8, Jan.

[125] New York City, "Hurricane evacuation map." Retrieved March 3, 2013, http://gis.nyc.gov/oem/he/map.htm, 2012.

[126] N. Andrienko, G. Andrienko, and P. Gatalsky, "Exploratory spatio-temporal visualization: an analytical review," *Journal of Visual Languages & Computing*, vol. 14, no. 6, pp. 503 – 541, 2003. Visual Data Mining.

[127] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97 – 136, 1980.

[128] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.

[129] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," in *ACM Sigmod Record*, vol. 28, pp. 49–60, ACM, 1999.

[130] J. Chen, M. K. Leung, and Y. Gao, "Noisy logo recognition using line segment hausdorff distance," *Pattern recognition*, vol. 36, pp. 943–955, 2003.

[131] J.-G. Lee, J. Han, and X. Li, "Trajectory outlier detection: A partition-and-detect framework," in *IEEE International Conference on Data Engineering*, pp. 140–149, IEEE, 2008.

[132] F. B. Viégas and M. Wattenberg, "Timelines tag clouds and the case for vernacular visualization," *interactions*, vol. 15, no. 4, pp. 49–52, 2008.

[133] M. Luboschik, H. Schumann, and H. Cords, "Particle-based labeling: Fast point-feature labeling without obscuring other visual features," *IEEE Trans. Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1237–1244, 2008.

[134] M. Dork, D. Gruen, C. Williamson, and S. Carpendale, "A visual backchannel for large-scale events," *IEEE Trans. Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1129–1138, 2010.

[135] F. Bergstrand and J. Landgren, "Visual reporting in time-critical work: Exploring video use in emergency response," in *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, (New York, NY, USA), pp. 415–424, ACM, 2011.

[136] Purdue University, "Engineering webcams." Retrieved December 3, 2014, https://engineering.purdue.edu/ECN/WebCam, 2014.

[137] V. J. Kok, M. K. Lim, and C. S. Chan, "Crowd behavior analysis: A review where physics meets biology," *Neurocomputing*, 2015.

[138] R. Maciejewski, R. Hafen, S. Rudolph, S. G. Larew, M. A. Mitchell, W. S. Cleveland, and D. S. Ebert, "Forecasting hotspotsa predictive analytics approach," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 4, pp. 440–453, 2011.

[139] A. Malik, R. Maciejewski, S. Towers, S. McCullough, and D. S. Ebert, "Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement," *IEEE Trans. on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1863–1872, 2014.

[140] P. L. Smith, "Splines as a useful and convenient statistical tool," *The American Statistician*, vol. 33, no. 2, pp. 57–62, 1979.

[141] J. Krüger, P. Kipfer, P. Kondratieva, and R. Westermann, "A particle system for interactive visualization of 3D flows," *IEEE Trans. on Visualization and Computer Graphics*, vol. 11, pp. 744–756, 2005.

[142] C. Beccario, "A project to visualize global weather conditions." Retrieved February 14, 2016, https://github.com/cambecc/earth, 2015.

[143] Kaggle, "Ecml/pkdd 15: Taxi trajectory prediction (i)." Retrieved September 1, 2016, https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i/data, 2016.

[144] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

VITA

VITA

Junghoon Chae is a PhD student in the School of Electrical and Computer Engineering at Purdue University. My research expertise and interests are, but not limited to, in the areas of visual analytics for large-scale data, spatiotemporal data modeling and visualization, and social media and text data mining. He received his Master of Science degree in Electrical and Computer Engineering from Purdue University in 2011 and Bachelor of Science degree in Computer Engineering and Electrical Engineering (Dual Major) from Kyung Hee University, South Korea in 2008. Contact him at jchae@purdue.edu.